FINAL REPORT

The MedSORT-II Project: Developing and Managing Medical Thesauri

Contract No. NOl-LM-6-3512

Principal Investigator:

David A. Evans

Departments of Philosophy and Computer Science
Carnegie Mellon University

Task Period:

July **1,** 1986 - August 30, 1987

Report Date:

August 28, 1987

FINAL REPORT
The MedSORT-II Project: Developing and Managing Medical Thesauri

Contract No. NO1-LM-6-3512

Principal Investigator:

David A. Evans

Carnegie Mellon University

Executive Summary. The MedSORT-II Project was conducted at Carnegie Mellon University over a twelve-month period, from July 1986 through June 1987, under a contract from the National Library of Medicine (No. NO1-LM-6-3512). The termination date of the Project was extended from June 30, 1987 to August 30, 1987, though the principal work was completed in the original twelve months. The Project Final Report has two main divisions: (1) a five-part essay focusing on *Project goals and accomplishments*, *the design of the Project thesaurus/knowledge base*, *experiments in natural-language processing*, *references*, and *a guide to the appendices*; and (2) nine volumes of supporting documentation (amounting to approximately 2000 pages), referenced by the essay.

The MedSORT-II Project has had two principal goals: (1) the development of a frame-based, semantic network for concepts in biomedicine (a thesaurus); and (2) the development of an enhanced knowledge-base management system. These goals have been accomplished in the form of a thesaurus/knowledge base of concepts structured around the notion of a *finding* in clinical, diagnostic medicine; and a frame-base-management system, SPAWN. In addition, the Project has sought both to validate the design decisions affecting the thesaurus and to demonstrate its utility by developing a natural-language interpretation facility based upon it. One goal has been to demonstrate the ability to translate natural-language statements of selected *clinical findings* into frame-representation structures reflecting the findings' semantic and pragmatic well-formedness conditions. Another has been to manage predictable variation in the expression of concepts, as that due to syntactic and morphological variation.

This report is divided into five parts. Part I summarizes the goals and accomplishments of the Project. Part II describes the development, organization, and management of the thesaurus/knowledge base, including a discussion of SPAWN. Part III presents the work on natural-language processing, including a description of the Project lexicon, the treatment of morphological variation, and the bottom-up parser and grammar used in our application demonstration. Part IV collects the references of the report. Part V contains a guide to the appendices—adumbrating the studies, sub-reports, data, copies of code, and other documents relevant to the Project. Conclusions and recommendations are reported in Parts I, and *interalia*, II, and III. Besides the Project Final Report, MedSORT-II Project deliverables include computer programs and computer-held databases. Portions of these have been reproduced in the appendices, but certain files, owing to their size and complexity, have been delivered to the National Library of Medicine in electronic form.

Contents

List of Figures

Part I
The MedSORT-II Project: Goals and Accomplishments

1. Introduction

The MedSORT-II Project was conducted exclusively at Carnegie Mellon University between July 1, 1986 and June 30, 1987 under support of a contract from the National Library of Medicine (No. NO1-LM-6-3512). The reporting period of the Project was formally extended to August 30, 1987, but all substantial work was completed by June 30, 1987. This report, consisting of a five-part essay and separate appendices, recounts the background considerations, the conceptual basis, the details, and the results of the Project's year-long effort.

The principal achievements of the Project are presented in Parts I, II, and III. Part I describes the Project background and offers summaries of the Project's goals and accomplishments. Part II reports on the central focus of Project activities—the development of a semantically coherent thesaurus/knowledge base (the MedSORT-II Thesaurus) and a knowledge-base management system (SPAWN). In addition, Part II describes the studies we have conducted to *validate* semantic classifications and semantic relations in our thesaurus. Finally, Part III discusses the natural-language processing applications we have developed as a means of evaluating the soundness and utility of our thesaurus design. In particular, we report on our morphological analysis program (MORPH) and on the modified bottom-up parsing system (derivative of RULEPAR) we have developed to process subsets of clinical findings. At various points in Parts I, II, and III, as well, we include conclusions and recommendations for future work.

2. Project Philosophy

At the most general level, the goal of the MedSORT-II Project has been to advance our ability to process natural-language textual data in the biomedical domain. Clearly, such an ability would have great implications for research, for health-care delivery, and for the management of medical information. Clearly, too, such capabilities cannot be achieved in the scope of a single project, but depend on the cumulative results of many efforts, each having perhaps highly specialized goals. It is in this perspective that the MedSORT-II Project should be judged. We have focused principally on the problem of developing and controlling thesauri/knowledge bases that could serve as the basis for processing the *concepts* designated by natural-language expressions. Thus, we have been concerned with semantic classifications and issues of semantic-network coherence appropriate to natural-language processing applications, without bias for particular styles of natural-language processing. Though we offer examples of approaches to natural-language processing, it is not our purpose to advocate a particular approach and it was not a goal of the Project to develop an extensive natural-language processing facility. Our chief objective has been to establish principles for the construction of semantically coherent thesauri, organized to reflect conceptual and linguistic-semantic relations among terms in the biomedical domain.

The MedSORT-II Project has developed in the context of several other activities involving groups in the Carnegie Mellon/University of Pittsburgh community and at the National Library of Medicine. These activities include the work on the INTERNIST-l/QMR System[1] and the MARS Project,[2] and, more recently, the joint efforts encompassed by the Unified Medical Language System (UMLS) Project.[3] Intellectual forces cannot propagate in a vacuum; the wave-front of scientific progress cannot be circumscribed by the artificial barriors of a project boundary: it is natural that the activities in such complementary efforts should influence one another. Thus, the genesis of the ideas that have found expression in the MedSORT-II Project cannot be traced to a single source; and the goals of our research have been modified freely in the course of twelve months to reflect refinements in our understanding of the central issues in developing biomedical-knowledge resources. Indeed, the special concerns of the MedSORT-II Project have been determined, in part, by a desire to contribute relevant resources to the common research community. But while the goals of MedSORT-II Project have been informed by the developing research in parallel projects, the Project philosophy and methodology has been influenced most directly by its genetic parent, the MedSORT-I Project. As a further basis for understanding the work of the MedSORT-II Project, then, we should consider in somewhat greater detail both the conclusions of the MedSORT-I Project and the specific statement of proposed activity that defined the objectives of our effort.

3. Background Context: MedSORT-I

The forerunner of the MedSORT-II Project was the MedSORT Project[4] conducted at Carnegie Mellon University under the support of the National Library of Medicine from September 1984 to December 1985. To distinguish that project from the one described in this report, we refer to it as the "MedSORT-f Project.

The MedSORT-I Project encompassed a number of activities, including biomedical-domain analysis, thesaurus construction, knowledge representation design and management, natural-language processing, study of indexing practices, and discourse analysis. The focus of each

[1] The INTERNIST-I/QMR System represents a refinement and application of the INTERMST-I Knowledge Base and System, and is under the direction of Dr. Randolph A. Miller, M.D. The System is based in the Section for Medical Informatics, Department of Medicine, School of Medicine, The University of Pittsburgh.

[2] The MARS Project is directed by Dr. John K. Vries, M.D., and is based in the Decision Systems Laboratory. School of Medicine, The University of Pittsburgh. The Project's goal is to develop a practical. automated medical-data archiving and retrieval system, with special test applications in the domain of Neurophysiology and Neuropathology. For additional information, see [Vries, *et al.* 1986].

[3] The UMLS Project is supported and directed by the National Library of Medicine. It currently involves several groups of researchers nation-wide and is in the first of two-years of exploratory activity. A principal goal of the Project is to develop methods and standards for the representation of biomedical information, especially for use in computational processing of textual data.

[4] "MedsORT" is a permuted acronym deriving from *Subject-Oriented Retrieval of Text in the Bio-Medical Domain.* The principal investigators of the MedsoRT Project were Jaime G. Carbonell, David A. Evans, Dana S. Scott, and Richmond H. Thomason.

activity was necessarily limited. Under *biomedical-domain analysis,* the Project considered only a subset of topics that are included under "Management and Treatment of Rheumatoid Arthritis". Under *thesaurus construction,* the Project developed a thesaurus (the MedSORT-I Thesaurus) sufficient to classify terms in the target domain, leading to detailed representations of some specific areas—such as *anatomy* and *rheumatologic diseases*—but lacking breadth and generality. Under *knowledge representation design and management,* it produced the frame-based-representation management system, FRAMEKIT. Under *natural-language processing,* it experimented in reapplications of two case-frame parsers, DVPAR-I and DYPAR-IV,[5] and developed RULEPAR/RULEKIT. Under *study of indexing practice,* it conducted protocol analyses of indexers at the National Library of Medicine. And under *discourse analysis,* it attempted to identify the conceptual, linguistic, and rhetorical structure of abstracts of articles in the domain under study. In almost every phase of activity, the Project was exploratory and creative, especially in attempting to bring principles from Artificial Intelligence and Cognitive Science to bear on the problem of representing natural-language textual information.[6]

The MedSORT-I Project reached several important conclusions. First, the Project emphasized the need for *semantically complex representations* of the structure of expressions used to drive indexing and retrieval. Second, the Project advocated the use of *natural-language principles* for developing generalized semantic networks. Third, the Project identified a *semantically rich thesaurus* as an appropriate resource in which to base representations. Fourth, the Project stressed the importance of *perspectives and tangled-hierarchical classification* in capturing relations in a domain such as biomedicine. And, finally, the Project recommended that an effort begin to develop an *Indexer's Workbench,* in part, to eliminate the inevitable errors introduced by indexers faced with the task of choosing appropriate labels for bibliographic records from among complex (and shifting) concept-terms. The strong, general recommendation of the Project was to exploit techniques in Artificial Intelligence, especially Knowledge Representation and Computational Linguistics, to achieve and manage the representational complexity required for processing *concepts*—not merely *terms*—in all their linguistic contexts.

The MedSORT-I Project, which had a budget approximately *three-times,* and a person-effort approximately *two-times* greater than the MedSORT-II Project, could not be renewed under the terms of its governing contract and was not continued in its original design.

4. MedSORT-II Goals

The MedSORT-II Project developed as a direct response to XLM RFP No. S6-103/PSP. Under the rubric of "Automated Classification and Retrieval," several broad areas of investigation were defined in that statement of goals, as noted in the following extracts:

[5] For information on the DYPAR family of parsers, see [Carbonell *k* Hayes 1981], [Carbonell & Hayes 1983], and [Carbonell, *et ah* 1983].

[6] A complete description of the MedsoRT-l Project can be found in [Carbonell, *et ai* 1985].

> The Lister Hill National Center for Biomedical Communications (LHNCBC) has developed a research program to investigate, develop, and evaluate computer science techniques which support the automated classification and retrieval of biomedical literature. ... The program will include projects in the areas of natural language understanding, knowledge representation, and information retrieval. The goal of these projects is to explore the application of these areas to the development of automated systems for identifying, representing, and retrieving concepts and main ideas from printed documents.
>
> ... The goal of the LHNCBC research in automated classification and retrieval is to improve the indexing and information retrieval processes for the enhancement of user access to the biomedical literature.
>
> Accordingly, the contractor shall ... conduct studies, analyses, and evaluations in selected areas of computer science directed at solving problems related to automated text processing and knowledge presentation.

In addition, the RFP specified the following foci:

> **Automated Merging of Existing Hierarchical Structures** LHNCBC is interested in the development of assessments and proposals as to how to build biomedical classification structures that can be used in the indexing and retrieval of biomedical documents. The long history of activity in biomedicine has witnessed the production of many manual and computerized codifications of information some of which might be usefully and easily transformed into the kind of information which would facilitate more intelligent, semi-automated processing of natural language material.
>
> **Designing Tools for Building Thesauri and Knowledge** Bases The contractor should examine biomedical thesauri and devise methods and tools for building thesauri that are more comprehensive in their coverage than existing thesauri. Existing thesauri and knowledge bases (a knowledge base is typically more extensive than a thesaurus) are based on some underlying knowledge representation scheme. The scheme, for thesauri, is usually very simple and involves only one or two binary relations. On the other hand, the scheme employed for knowledge bases is quite extensive and involves many relations and inference rules. The contractor should investigate how existing biomedical thesauri, including MeSH, SNOMED. and ICDA, may be extended to knowledge bases and may be used, to extend and maintain existing knowledge bases.

These goals are consistent with the spirit of MedSORT-II research and with the more specific objectives we set for ourselves, though we have not addressed every facet of activity cited above. Our actual program of research was determined in outline form by the Project Proposal that set forth our principal goals.

4.1. Project Proposal

Under the title "Developing and Managing Medical Thesauri," the MedSORT-II Project proposed specifically to focus on two important aspects of the indexing and retrieval problem, given the background conclusions of the MedSORT-I Project: (1) thesaurus development (amounting to 70% of the proposed effort); and (2) FRAMEKIT enhancements (amounting to 30% of the proposed effort). In particular, under thesaurus development, we proposed to revise and expand the MedSORT-I Thesaurus to reflect concepts in general medicine (as exemplified in the INTERNIST-I/QMR knowledge base), in Neurology and Neuropathology (as defined by the work of our colleagues at the Decision Systems Laboratory at the University of Pittsburgh), and in MeSH. Under FRAMEKIT enhancements, we proposed to develop facilities for the control of *perspectives* and for the inheritance of features via non-*isa* relations.[7]

In practice, we have had to specify further the interpretation of these goals. A mindless expansion of the MedSORT-I Thesaurus would serve no purpose, for example. Thus, we identified an idealized application—the processing of natural-language statements for expressions characterizing clinical findings—to sharpen our focus. In fact, this application served to guide not only the construction of an appropriate biomedical thesaurus, but also the development of the enhanced frame-management system (SPAWN) that represents our extension of the MedSORT-I-Project system, FRAMEKIT. It also guided us in the development of natural-language processing software for use in validating the thesaurus and frame-management system and for evaluating our work.

4.2. Project Deliverables

Besides this Final Report, Project deliverables have included a 6-Month Oral Report[8] and copies, with suitable documentation, of the databases and software developed in the context of the Project. These are principally the MedSORT-II Thesaurus and the enhanced knowledge-representation and management system, SPAWN. Though there was no specific obligation to develop natural-language processing software, the MedSORT-II Project has also included among its deliverables copies of (1) a modified version of RULEPAR, RULEPAR-II, and (2) the Project morphology analyzer, MORPH, along with associated resources.

Finally, in addition to the written documentation represented by this Final Report with accompanying Appendices and the copies of software and knowledge-base resources, we have delivered on magnetic tape a virtual image of all the programs and databases we have developed for the evaluation of the thesaurus/knowledge base and for natural-language processing experiments. In effect, it will be possible for the National Library of Medicine staff to run our programs directly on data we have used. To facilitate transportability, we have delivered all software in FRANZLISP.

5. MedSORT-II Accomplishments

The following summary highlights the accomplishments of the MedSORT-II Project: details are recounted in Parts II and III of this report. Accomplishments can be categorized broadly under (1) *thesaurus/knowledge base development*; (2) *knowledge management*; and (3) *natural-language processing.*

[7] The complete statement of our proposed MedSORT-II Project activity is included, in its original form, in Appendix A1.

[8] The 6-Month Report was delivered in a presentation to the National Library of Medicine on 16 January 1987. A copy of the overhead-projection transparencies used in that presentation is included as Appendix A2.

5.1. Accomplishments in Thesaurus/Knowledge Base Development

The concrete manifestation of MedSORT-II Project work on developing a thesaurus/knowledge base suitable to the representation of concepts in the biomedical domain is the MedSORT-II Thesaurus. It contains representations of more than 5700 canonicalized, *basic* concepts from *clinical, diagnostic medicine;* more than 100 *complex* concept-clusters (including representations for *laboratory findings);* and links to MeSH categories A, B, and C. The concepts are embedded in tangled, hierarchical networks (defined principally by *>a-relational links); and a subset of the concepts (chiefly from *anatomy)* have *non-isa*-relational links to one another (including the relations *part-of* and *contains* and their inverses). Because the concepts represented are both canonicalized and basic, the thesaurus actually encompasses a large subset of possible natural-language expressions in the domain. First, when used with a lexicon and morphological-analysis software (such as described briefly, below), it can capture many times its concept-number of terms in their natural-language-variant forms. Second, when used in conjunction with suitable parsers (such as the modified version of RULEPAR we have produced), it can capture a virtually unlimited number of *complex* concepts—typically the syntactic variants of basic-concept combinations.

The MedSORT-II Project Thesaurus has been created *de novo,* with its core structure and content based on a study of the set of findings in the INTERNIST-l/QMR knowledge base. In addition, the thesaurus has been supplemented by the addition of terms from the MedSORT-I Thesaurus and terms derived from a study of the MARS Thesaurus.

While the Thesaurus can stand alone as a taxonomic classification of basic terms in clinical medicine, its special utility derives from its design with natural-language processing applications in mind. The Thesaurus is not only *semantically coherent,* but is designed to capture *pragmatic contexts* naturally, allowing one to define parallel taxonomies of concepts of arbitrary complexity, composed of the basic concepts it includes.

In addition, project work on the thesaurus/knowledge base has resulted in the development of a methodology for validating semantic relations among concepts in domains such as biomedicine.

5.2. Accomplishments in Knowledge Management

Our ability to capture concepts comes not only from their identification and organization in a thesaurus, but also from the management of the data structures that encode them. Rather than modify the MedSORT-I Project frame-management system. FRAMEKIT, we have designed an enhanced, new system, SPAWN. SPAWN reproduces all the functionality of FRAMEKIT but goes beyond it to include perspective management (referred to as *views)* and arbitrary definition of inheritance relations. While SPAWN has been especially useful in implementing the structure of the MedSORT-II Project Thesaurus, it is a general frame-base management system and can be used to implement more traditional thesauri, such as MeSH.

Along with SPAWN, we have produced a user's manual and scripted examples of user-SPAWN interaction.

5.3. Accomplishments in Natural-Language Processing

In addition to the thesaurus/knowledge base and SPAWN, the MedSORT-II Project has developed a number of utilities to facilitate use of the thesaurus in natural-language processing applications. These include a lexion, a morphological analysis program capable of recognizing syntactic variation due to selected inflectional morphology, MORPH, and a modified (and extended) version of the MedSORT-I Project parser, RULEPAR-II. The parser, in particular, is capable of returning analyses of *partial* information and is sensitive to (and driven by) pragmatic contexts—nż., just those combinations of concepts that have a special value in the epistemological structure of the domain of *clinical medicine.*

Each of the programs can stand alone or be used in other applications than the ones we have chosen. Thus, the natural-language processing software can be regarded as independent and general facilities for use in managing natural-language data.

6. Conclusions **and Recommendations**

We note our conclusions and recommendations on specific Project activities at various points in Parts II and III. Here, we offer thoughts on the general conclusions and recommendations to be drawn from our experience.

6.1. General Project Conclusions

The achievements of the MedSORT-II Project suggest five general conclusions, with implications for the development of future generations of automated indexing and retrieval systems.

1. It is possible to develop semantically coherent thesauri/knowledge bases that go beyond the limitations of traditional thesauri (which cannot be used effectively in natural-language processing applications) without having to develop fully-instantiated knowledge bases. Such resources can be organized into frame-based schemata by (1) developing a taxonomy of *basic* concepts following linguistic-semantic principles; (2) defining parallel taxonomies of selected combinations of basic concepts following domain-pragmatic principles; and (3) encoding a limited number of non-?s<z-relational links to provide a basis for enforcing semantic-selectional restrictions among concept combinations.

2. Thesauri of more conventional design can be merged into networks such as those represented by the Project thesaurus/knowledge base. Large, multiply-hierarchical thesauri, such as MeSH, can be captured *in toto* as parallel *perspectlves*, making contact with the core, basic-concepts hierarchy at just those concept-nodes shared by each separate thesaurus. Smaller, specialized thesauri, such as the MedSORT-I or MARS Thesaurus, can be added by developing sub-classifications of basic concepts and representing domain-specific combinations (the *narrower-than* terms) as complex concepts in a parallel hierarchy.

3. The important notions of clinical medicine, including *clinical finding* and *laboratory procedure,* for example, are semantically and pragmatically well-formed. There are good examples of idealized contexts and concept-combinations in the knowledge bases of expert systems such as INTERNIST-l/QMR and HELP. Thus, it is possible to develop *grammars* for sub-domains of biomedicine that will not only aid the design of the thesauri/knowledge bases that organize the concepts, but will also provide direct links to the natural-language structures that give them textual expression.

4. Relatively stream-lined knowledge-management systems are adequate for the implementation of thesauri/knowledge bases. As minimum facilities, they should be able to support (1) general inheritance, (2) arbitrary definition of relations, (3) some form of perspective management, and (4) unlimited tangling.

5. In well-structured domains such as biomedicine, natural-language interpretation can be driven by semantic-functional parsing. In particular, it is not necessary to have fully-developed natural-language understanding to generate representations of expressions in natural language that capture the contexts and relations among concepts required for applications in accurate indexing.

More practically, our experiences cause us to reflect on the problems of developing and executing a sustained, coherent research plan, with goals such as those undertaken by the Project. Some of these thoughts resonate with conclusions of the MedSORT-I Project. We identify the most urgent of these observations below:

6. As in the case of MedSORT-I work, we found that most of our effort—in time and intellectual energy—has been devoted to the task of organizing, developing, and refining the thesaurus/knowledge base. *Classification takes time; good classification takes more time.* Any project attempting to develop semantic taxonomies, especially in frame-based systems, must plan on numerous iterations of classification activity.

7. Domain experts must be involved in identifying the pragmatic contexts and, to a lesser extent, the semantic classes to be included in a thesaurus/knowledge base such as the one we developed. In our case, experts in clinical medicine were essential to the classification task, though principally as critics of the *Praxis,* not the *Theorie,* of our approach. Projects that seek to combine computation, linguistics, and concepts from clinical medicine *are* interdisciplinary; their teams must be interdisciplinary, as well.

8. Though only a one-year effort and limited in scope, the MedSORT-II Project has many accomplishments. Nevertheless, we firmly believe that short-term research contracts are not an appropriate vehicle for work on problems of the scope we confronted. If we are to hope for comprehensive solutions to the automated classification and retrieval goals articulated by the Lister Hill National Center for Biomedical Communications, we must be willing to allocate more resources and longer periods of time to the teams of workers conducting the basic research. For research-university-based groups in particular, the discontinuity of short-term project support is especially destructive and counterproductive of the goal of building necessary, stable research teams.

The tenor of these conclusions should be clear: the fundamental principles of a computational-linguistic approach to the development of biomedical thesauri are sound and the MedSORT-II Project Thesaurus and knowledge-management system can serve as prototypes for future work; but the development of general computational-linguistic resources for use in the biomedical domain will require considerable time, effort, and support.

6.2. General Project Recommendations

Many of the conclusions noted above suggest natural extensions of work or modifications of policy. Rather than belabor the obvious reflexes of the Project's conclusions, we make only two specific recommendations for future activities: to focus attention on the expansion of the MedSORT-II thesaurus/knowledge base and to merge the efforts begun under MedSORT-II in both knowledge representation and natural-language processing into the activities of longer-term projects, such as the UMLS Project. We elaborate on these points briefly below.

1. The MedSORT-II thesaurus/knowledge base is a potentially powerful resource for both biomedical-knowledge representation and natural-language processing. *It should be developed further.* As demonstrated by our experiments in natural-language processing, it can be used in its present form to support the identification of a subset of clinical findings. With additional extensions—at the level of concept clusters—it could support practical applications, such as automated indexing of hospital charts.

2. An obvious solution to the problem of limited time and resources experienced by the MedSORT-II Project is to link its continuing efforts to those of other projects. In particular, because of the Project's focus on issues directly related to developing canonical knowledge representation, ***we would advocate pursuing* MedSORT-II *goals under the* UMLS *Project***, in collaboration with colleagues participating in that project. In retrospect, we can observe that many of the desiderata of the MedSORT-I Project have been met in the design—and performance—of the MedSORT-II system. The combined thesaurus/knowledge base and the natural-language processing software, including parser, morphological analyzer, and lexicon, essentially realize in prototype form one of the principal objectives identified by the MedSORT-I Project for future work: *standardizing biomedical knowledge bases.*[9] Such standardization, it was argued, would involve (1) standardizing biomedical knowledge representation, (2) linking natural language to knowledge representation, (3) creation of a biomedical thesaurus, and (4) structuring biomedical information both cross-modally and cross-functionally. These are among the objectives of the UMLS Project, as well. We would claim that the MedSORT-II Project has achieved these goals for a coherent subset of clinical medicine and that it has demonstrated the means by which these goals might be realized for the general biomedical domain.

We trust that the conclusions and recommendations collected here are substantiated by the details of the Project activities, presented in Parts II and III of this report.

[9] *Cf.* pp. 34-35 of [Carbonell, *et al.* 1985] for a complete discussion of this objective.

7. Acknowledgements

Besides the Principal Investigator, the MedSORT-II Project team has included a number of researchers, consultants (from both the University of Pittsburgh and Carnegie Mellon University), and assistants. Over the course of the project year, the MedSORT-II Project team has consisted of the following individuals:

- *Research Associates*—Dr. Sandra Katz, D.A.; Dr. Lori Taft, Ph.D.
- *Research Programmers*—John Aronis; William Lott; Steve Morrisson
- *Graduate Student Assistants*—Armar Archbold; Edward Gibson; Rick Kazman; Jinhao Wang; Philip Werner
- *Undergraduate Student Assistants*—Andi Blaustein; Sam Hennessey; Thomas Kuhn; Thuy Nguyen; Matt Nolan
- *Consultants*—Dr. Gordon Banks, M.D., Ph.D. (Pitt); Dr. Jaime Carbonell, Ph.D. (CMU); Dr. Nunzia Giuse, M.D. (Pitt); Dr. Fred E. Masarie, M.D. (Pitt); Dr. Thomas Medsger, M.D. (Pitt); Dr. Randolph A. Miller, M.D. (Pitt); Dr. Dana S. Scott, Ph.D. (CMU); Dr. Richmond H. Thomason, Ph.D. (Pitt); Dr. John K. Vries, M.D. (Pitt)
- *Clerical Staff*-—Kathryn Gula; Kristen Zocco

In addition, the MedSORT-II Project has received support and direction from Jules Aronson, the Project Officer at the National Library of Medicine, and Dr. Alexa McCray, Ph.D., the Alternate Project Officer.

The Project has benefited greatly from access to the facilities of the Laboratory for Computational Linguistics (LCL), in the Department of Philosophy, and also the facilities of the Department of Computer Science at Carnegie Mellon University. The Project Contract covered no costs for computation or equipment, so all the Project's computational activities were supported by these departments. The Project also benefited greatly from the generous cooperation of Dr. Randolph A. Miller, M.D., and the Section for Medical Informatics at the University of Pittsburgh School of Medicine. In particular, our work would have been impossible had we not had access to the INTERNIST-l/QMR knowledge base, which served as our principal resource in developing an idealized conception of medical findings.[10] We wish to acknowledge as well the warm cooperation of Dr. John K. Vries, M.D., who allowed the Project to analyze the terms and structure of the MARS Thesaurus.

We owe a special thanks to Dr. Fred E. Masarie, M.D., and Dr. Nunzia Giuse, M.D., whose timely advice and constructive criticism more than once saved us from misguided efforts. We appreciate especially their patience and their willingness to consider the problems of representing concepts in the biomedical domain from our point of view, often at odds with conventional medical sensibilities.

Almost every member of the MedSORT-II Project team contributed directly to the drafting of this report and the compilation of appendices. In preparing this document, individual

[10] **Our access to the INTERNIST-I/QMR knowledge base came in the form of data files, the QMR System, and programs for browsing INTERNIST-I/QMR data developed in the Section for Medical Informatics.**

team members, besides the Principal Investigator, were most directly responsible for working on the following sections as indicated:

- *Knowledge Development*—Sandra Katz, Thomas Kuhn
- ***Knowledge Validation*—Armar Archbold**
- *Knowledge Management*—Rick Kazman, Andi Blaustein
- *Linguistic Variation*—Lori Taft, Jinhao Wang
- ***Syntactic Processing*—Edward Gibson**

Much that is good in this report derives directly from their hard work and dedication. None of the members of the Project team bears any responsibility for its shortcomings except the Principal Investigator.

Part II

A Thesaurus for Concepts in Clinical Medicine

1. Introduction

The chief work of the MedSORT-II Project has involved the development of the MedSORT-II Thesaurus and SPAWN. This portion of the Final Report is devoted to a description of that work. In Section 2, *Thesaurus Development,* we discuss the objectives, procedures, experiences, and results of our work in the development of the MedSORT-II Thesaurus. We describe especially the construction and extension of the core, *basic-concept* classification hierarchy and the development of *complex-concept* clusters appropriate to *clinical findings.* In Section 3, *Exploring Semantic Classes and Relations,* we recount some of the studies we conducted in the course of the development of the thesaurus/knowledge base. In particular, we report the results of our analyses of findings phrases from the INTERNIST-l/QMR knowledge base and our attempts both to discover and to validate semantic relations among concepts, as revealed by the co-occurrence distributions of terms in findings phrases. Finally, in Section 4, *Managing Representations of Knowledge,* we describe the principles of frame-based representation and our specific implementation of them in the Project knowledge-management system, SPAWN.

1.1. Background Considerations in Project Thesaurus Design

Much of what we describe in our work assumes an approach to problems in knowledge representation that derives from research in Artificial Intelligence, Computational Linguistics, and more particularly, the experiences of the MedSORT-I Project.[11] One basic premise is that knowledge can be analyzed as composed of atomic units—concepts, including relations—whose properties can be specified explicitly and formally and whose interpretation can be determined independent of context. Another basic premise is that structured representations can be used to capture the semantic and pragmatic relations that characterize constellations of concepts; and that the salient relations are both general *(i.e.,* non-concept-specific) and limited in number.

While it may be possible to represent concepts context-independent!}-, applications of concept-representations to specific tasks may impose further constraints that are best mirrored in context-sensitive organizations of concepts. Thus, our special approach—guided by the focus we have taken on *natural-language processing*—reflects the need we have to link concepts to natural-language objects and to be able to capture generalizations over classes of concepts that correspond to the semantic selectional restrictions characteristic of rules in generative grammars. The thesaurus/knowledge base we have designed, therefore,

[11] **For references on the background literature in this area, see [Brachman *fz* Levesque 1985].**

has a special utility in developing representations for concepts that will map readily onto natural-language lexical items and phrases; but it does not lose its validity as an independent resource for concept classification, similar to more conventional thesauri.

The specific approach to categorization and representation of concepts we have taken derives from the experiences of the Carnegie Mellon University research effort in the UMLS Project. The principal objective of that work has been to characterize biomedical concepts in parallel taxonomies of increasing complexity: *basic* concepts define the atoms that compose *complex* concepts; and complex *concept-clusters* represent privileged contexts for interpretation. In practice, three parallel hierarchies of concepts can be used to organize information required for representing *clinical findings*—a basic (Bx) taxonomy, a concept-cluster (Cx) taxonomy, and a findings-concept-cluster (Mx) taxonomy. The approach is designed to make explicit the pragmatic-context sensitivity that resides at the heart of domain-specific expertise.[12]

In developing our thesaurus/knowledge base, we have found it important to keep the distinction between *concepts* and *words* clearly in mind. When working with *terminologies,* the two can easily become confused. Concepts provide the epistemological structure of a domain. Some concepts are *effable;* these frequently become associated with specific lexical items—bound-morphemes, free morphemes (words), and phrases. Besides the truely *ineffable* concepts (such as imagistic concepts), there are many non-lexical concepts trapped in the network of relations created by lexical concepts, which are not readily expressible as lexical items or phrases. In practice, such concepts may manifest themselves as the *well-formedness* conditions on clusters of lexical items in phrases.[13] In developing a thesaurus/knowledge base, we must design structures that can *capture* concepts; and not merely develop a latticework of lexical items. We believe we have found in the structure of the MedSORT-II Thesaurus an appropriate means for doing this in the biomedical domain.

1.2. Orientation to the Project Effort

Our goal was to develop a *prototype* thesaurus for concepts in clinical medicine. Hence, we have not aimed to produce an exhaustive and fully-structured network for all of biomedicine or even all of clinical medicine. Rather, our purpose has been to produce a structurally-sound and selectively-elaborated network to demonstrate the feasibility of our design. This qualification, however, should not be taken to suggest that the thesaurus is not *complete* or *comprehensive.* In fact, we would claim that virtually all the expressions associated with findings recognized by the INTERNIST-l/QMR system can be expressed by concepts represented in the Project Thesaurus.

We should note that the Project Thesaurus actually represents several resources, including a set of conventions for standardizing string-forms, a lexicon, a system for handling morphological variation, and a body of implicit pragmatic knowledge, in the form of grammatical

[12] A portion of the report on the UMLS effort in which this approach is elaborated is included as Appendix Bl.

[13] See [Evans *k,* Scott 1986] for further discussion of these points.

constructs specifying the structure of concept-clusters. In this respect, the Thesaurus is more like a knowledge base than a conventional thesaurus, so we are wont to refer to it as a *thesaurus/knowledge base*. Many of the activities we have engaged in—as reported in the following sections—have been addressed to questions that arise in this broader perspective. Of course, we have attempted to focus our discussion by describing *components* of the system individually. Where clear divisions of representational responsibility exist, as in the handling of morphological variation, for example, this presents no problem; but the boundaries are not always so sharp.[14]

We should point out, as well, that our work has necessarily involved parallel activities—one result of which is that we have multiple resources in different stages of completion. The studies we report have validity for the *core* thesaurus/knowledge base; but may not reflect the latest developments, especially as required for our demonstration of natural-language processing, which demanded refinements in the representations of concepts that went beyond those reflected in the thesaurus/knowledge base, in some respects.

[14] Aspects of the Project most clearly associated with natural-language processing concerns, including the development of MORPH and the demonstration of processing of clinical findings, are presented in Part III.

2. Thesaurus Construction

The development of large-scale thesauri primarily involves the identification and classification of significant concepts, analysis of their semantic structure, and careful checking to ensure that the thesaurus is accurate and consistent. This section represents a report on the procedures involved in building the MedSORT-II Project Thesaurus, validating it, and integrating it with other components of the system.

2.1. Background Objectives

Our principal objective was to build a semantically coherent thesaurus/knowledge base that would enable us to combine expressions into meaningful phrases—in our case, a candidate statement of a finding in clinical medicine. For example, we would expect our thesaurus to support the mapping of the following clinically equivalent expressions in natural language,

- *chest pain made worse by swallowing,*
- *substernal pain worse with swallowing,* and
- *chest hurts, especially when patient eats,*

to a canonical representation of the appropriate finding, as exemplified by the phrase, CHEST PAIN SUBSTERNAL EXACERBATION WITH SWALLOWING, as given in the INTERNIST-l/QMR knowledge base.

To determine appropriate semantic structures for medical findings, we analyzed the approximately 4100 finding phrases contained in the INTERNIST-l/QMR knowledge base. We chose to use the INTERNIST-l/QMR knowledge base both because of its availability and also because it represents an extraordinarily complete and consistent source of findings expressions in the domain of diagnostic internal medicine. A complete analysis entailed identifying the basic concepts that combine to form finding expressions and analyzing the semantic relationships between basic concepts.[15] For example, five principal concepts comprise the complex finding phrase, CHEST PAIN SUBSTERNAL EXACERBATION WITH SWALLOWING: *chest, pain, substernal, exacerbation, swallowing.* Analyzing this phrase, we see that *pain* is what the patient reported to the physician, *chest* and *substernal* are locative modifiers, and *with swallowing* provides a conditional context for the event of *pain.* By comparing this expression to other expressions in the INTERNIST-l/QMR database, we can identify classes of *reports, locations, contexts, etc.,* that serve to define functional semantic categories for concepts.

Another of our principal objectives was to integrate other sources of medical information into the conceptual framework that we designed and implemented by analyzing the INTERNIST-l/QMR findings. We merged concepts from three sources: the thesaurus constructed during the MedSORT-I Project, the neuropathology lexicon under development by the MARS

[15] **This work followed the direction of concept classification and identification undertaken during Task 2 of the Unified Medical Language System (UMLS) Project at Carnegie Mellon University. For reference, a** copy **of the relevant portion of that report is included as Appendix Bl.**

Project [Vries, *et al.* 1986], and MeSH.

2.2. Summary of Procedure and Rationale

The frame-based MedSORT-II Thesaurus can be described as a *hierarchically ordered, semantic network.* One advantage of this design is that it allows for *inheritance—i.e.,* lower-level frames automatically 'inherit' information encoded in higher-level frames that belong to the same branch of the classification hierarchy—thus facilitating concise, efficient representation of information. The key phrase in the preceding is *classification hierarchy]* each concept must be assigned to a semantic category based on its features, and each category must, in turn, be organized in an acyclic, directed network, with categories relatively closer to individual concepts being regarded as more *specific,* categories relatively further away being regarded as more *general,* concept types.[16]

Our first task was to create a complete listing of the basic medical concepts that comprise findings. Using the INTERNIST-l/QMR knowledge base as a first source, we identified approximately 4000 basic concepts. This was not a straightforward matter of finding every term that appears at least once in the INTERNIST-l/QMR findings, for some terms are nearly devoid of meaning on their own. For example, the term *Babinski* is medically significant only when followed by *sign; converting* is vacuous in the medical domain apart from *angiotensin-converting-enzyme; etc.* For this reason, we attempted to identify concepts in appropriate contexts, by referring to a listing of INTERNIST-l/QMR terms and their associated findings, a sample of which can be seen in Appendix B2.

Our next major task was to classify each concept in terms of the categories from the basic hierarchy developed under the UMLS Project, as shown in Appendix B3. For this task we used both standard medical dictionaries, such as *Dorland's Medical Dictionary* ([Dorland 1985]) and *Taber's Cyclopedic Medical Dictionary* ([Taber 1981]), and also the Project's medical consultants from the University of Pittsburgh. We wanted to ensure that concepts were not only assigned to appropriate categories in the hierarchy, but that principles of classification had been applied consistently. For example, we would not allow some bones to be listed under *macro-body-structure,* while others were designated *macro-body-pari.* To facilitate checking, we indexed our terms according to their category, as seen in the sample set of classified terms in Appendix B4. Our current version of the top portion of the hierarchy is given in Figure 1. Further details can be found in Figures 5-12, in later sections of the report.

Another important task involved developing cross-classification *(non-isa).* semantic relations. This was necessary not only to render the thesaurus serviceable in any serious natural-language processing applications, but also to demonstrate the soundness of the frame-managment system, SPAWN, in which the thesaurus is implemented. Chief among the semantic relations we identified is *containment*—in particular, the relation between an anatomical object and its constituents. INTERNIST-l/QMR proved to be a good source of information about containment relations among medical concepts, for terms designating gross

[16] **It is customary to refer to the more general concept types as *higher-level* categories.**

- Bx-Thing
    - **physical-thing**
        - **living-thing**
            - **organism**
        - **non-1iving-thing**
            - **body-thing**
            - **place**
            - **substance**
            - **instrument**
            - **pathological-factor**
    - **abstract-thing**
        - **meta-language**
        - **marker**
            - **relation**
            - **grammar**
                - **grammatical-marker**
                - **morphological-item**
                    - **bound-morpheme**
                    - **free-morpheme**
                        - **lexical-item**
                        - **phrasal-lexical-item**
        - **medical-procedure**
        - **circumstance**
            - **patient-circumstance**
            - physiological-circumstance
        - action/event
        - experience
            - patient-experience
            - behavior
        - measure-theoretic-thing
            - measure
            - unit
            - **quality**
            - **relative-index**
        - organizational-entity
            - discipline

Figure 1: Top Portion of the Bx Hierarchy

anatomical regions often co-occur with those designating more specific regions or body parts, and the relation between terms in such cases is almost always implicitly one of containment, as the following findings demonstrate:

- **ABDOMEN PAIN HYPOGASTRIUM**
- **CHEST BRUIT CONTINUOUS INTERSCAPULAR**
- **EAR <S> CALCIFICATION AURICULAR CARTILAGE BILATERAL**

The *abdomen* contains the *hypogastrium;* the *interscapular* region is part of the *chest;* and the *ear* contains the *auricular cartilage*—all play a role in locating the salient observation, *pain, bruit* and *calcification,* respectively.

By encoding containment relations in the thesaurus, we facilitate the automatic inference of a general anatomical structure given only elements representing specific anatomical parts, or *vice versa.* Without such information, the thesaurus could not direct us to consider appropriate generalizations (categories in the thesaurus) as possible interpretations of input expressions. For example, we might fail to associate an expression of an ailment involving the patient's stomach to generic categories for abdominal findings. Furthermore, since specific anatomical parts often share features with their 'containing' counterpart, inheritance can take place across containment relations, just as it can across classificatory relations. For these reasons, where appropriate, we encoded the relations *contains* and *part-of* and their inverses in the thesaurus.

After classifying basic concepts, we focused our attention on analyzing the semantic structure of findings. Again, following the design of the UMLS study on classification of concepts, we focused on the structure of a manifestation (Mx) hierarchy—i.e., a hierarchical ordering of findings, as given in Figure 2. At the uppermost levels, we adopted the INTERNIST-l/QMR database's approach of using *diagnostic method* as the governing classification principle. Thus, for example, we divided the findings into two major classes—those obtained by patient report *(Report-Mx)* and those obtained through the physician's observations *(Observation-Mx).* We subdivided the former into reports of medical history *(Patient-Histonj-Mx)* and reports of present symptoms *(Patient-Symptom-Mx):* we subdivided the latter into those observations obtained by physical examination *(Physical-Exam-Mx),* and those obtained through laboratory techniques. The latter was subdivided further, as shown in Appendix Bõ. At lower hierarchical levels, we applied the principle of saliency—*i.e.,* the observation that there is a central concept in each finding—to classify findings. For *report* and *observation* findings, the central concept is usually a *pathological factor* or an *evaluated attribute,* as illustrated in the following finding phrases (where the central concept is underlined):

- **DYSPNEA AT REST**
- **HAND PAIN ELICITED BY SUSTAINED FLEXION OF WRIST**
- **TASTE METALLIC**

In laboratory observation findings, the central concept and organizing principle is the specific laboratory technique applied:

```
Mx-Thing
        Report-Mx
                Patient-Hx-Mx
                        Patient-Demographic-Hx-Mx
                        Patient-Social-Hx-Mx
                        Patient-Medical-Hx-Mx
                        Exposure-Hx-Mx
                        Behavior-Hx-Mx
                        Ingestion-Hx-Mx
                        Event-Hx-Mx

                Patient-Symptom-Mx

        Observation-Mx
                Physician-Observation-Mx
                Lab-Source-Observation-Mx
                        Lab-Instrument-Mx
                                Lab-Monitoring-Mx
                                Laboratory-Test-Mx
                                Extraction-Mx
                                       Tissue-Extraction-Mx
                                       Fluid-Extraction-Mx
                                Test-Of-Dynamic-Function-Mx
                        Lab-Assay-Mx
                                Substance-Technique-Mx
                                Tissue-Technique-Mx
                        Lab-Imaging-Mx
                                Lab-Indirect-Imaging-Mx
                                Lab-Direct-Imaging-Mx
```

Figure 2: Principal Structure of the Mx Hierarchy

- ABDOMEN COMPUTERIZED TOMOGRAPHY SMALL INTESTINE INTRAMURAL GAS
- KIDNEY <S> ARTERIOGRAPHY ABERRANT RENAL ARTERY OBSTRUCTING URETERO-PELVIC JUNCTION
- SKIN LESION ACTINOMYCES BY STAIN

Once the general hierarchical scheme for manifestations was in place and each finding was assigned to an appropriate place in this system, we determined the semantic structure of concepts at each hierarchical level. We began with *Patient-History-Mx, Patient-Symptom-Mx, Physical-Exam-Mx*, and *Laboratory-Observation-Mx* by attempting to isolate general classes of *slots* and corresponding *fillers* for findings of each type. We approached these questions empirically—by analyzing sets of findings that correspond to each type of manifestation cited above, designing a frame that potentially captures the semantic structure of the findings, and revising frames as additional findings were analyzed. This led naturally to the identification of *concept clusters* (Cx) composed of categories from the UMLS basic hierarchy, as described in the discussion of concept organization reported in Appendix B1. For example, the broad

```
<EVALUATED-ATTRIBUTE-CX>
      Focus: restricted to be of Bx type <EVALUATED-ATTRIBUTE>
      Value: restricted to be of type <VALUE-CX>
```

Figure 3: Structure of *Evaluated-Attribute*

semantic structure of an *evaluated attribute* can be given as in Figure 3.

We have implemented our thesaurus as a frame-based semantic network in SPAWN,[17] which has facilities for ensuring consistency. For example, if a concept is referenced for which there is no corresponding frame, SPAWN will automatically create one. SPAWN'S ability to utilize *inverse* links also facilitates the semi-automatic building of knowledge bases and frees the user from the burdens of handcrafting individual frames and networks. Furthermore, SPAWN permits automatic inheritance of values from higher-level frames, and combines frames for the same concept into one frame. In contrast to checking consistency, checking the accuracy of the thesaurus is a much more difficult task, one that is continuing with the aid of medical consultants.

Given the basic semantic structure of the thesaurus as determined by concepts derived from INTERNIST-l/QMR, we have attempted to integrate other systems of medical-concept organization into the thesaurus. This has involved (1) classifying concepts from other sources in terms of our existing categories, particularly those from the UMLS basic hierarchy, and (2) extending the hierarchy to include more specific subcategories from the added sources; and using the *views* mechanism built into SPAWN.

The first method was applied to concepts from the MedSORT-I thesaurus and the MARS lexicon. We found that most of the concepts from these sources could be 'slotted' into our basic hierarchy. But each database presented us with particular challenges. For example, the MedSORT-I thesaurus contains a far broader coverage of anatomy than the one that we built upon INTERNIST-l/QMR concepts; in particular, it represents most articular and skeletal structures. In order to capture this detail in our basic classification hierarchy, we added two subcategories to *macro-body-structure: articular-macro-body-structure* and *skeletal-macro-body-structure.* Finally, the MedSORT-I thesaurus contained a more detailed hierarchical structure of *medical procedures,* both diagnostic and therapeutic, which we could readily augment our hierarchy to include. Appendix B6. shows the Bx hierarchy, expanded to accommodate the MedSORT-I thesaurus.

One positive side-effect of merging the MedSORT-I concepts into our thesaurus was a significant improvement in the classificatory structure of the former. Some of the terms that we could not classify adequately with the scheme developed during the the MedSORT-I project could now be assigned to more specific categories. For example, whereas *pulmonary-function-test* had previously been classified as a *lab-method,* we could use our more detailed sub-

[17] SPAWN is described in Section 4 of this Part.

classification of laboratory methods to reclassify this concept as a *test-of-dynamic-function.* Gross anatomical parts such as *head* and *trunk,* which had previously been classified as a *solid*, were reclassified as *macro-body-part, etc.*

One factor that made merging the MedSORT-I thesaurus into ours relatively straightforward was the fact that it implemented a well-developed, hierarchical semantic structure. This is not the case w'ith the MARS lexicon. The most well-developed aspects of the MARS database are the *broader-than* and *narrower-than* relations among concepts. Concepts in the lexicon, however, are not developed under semantic classes compatible with our thesaurus. For example, although each lexical entry is also assigned to a category called a *pathology-class*—intended as a taxonomic category—some of the class names are isomorphic to the lexical entry itself; others are grammatical categories *(e.g., noun, adjective, etc.)* rather than semantic classes. Consequently, it is much more difficult to assign categories from the UMLS basic hierarchy to the approximately 10,000 terms in the MARS database. We are currently reclassifying terms from the MARS database with the help of an expert in neuropathology.

We have relied on the *views* mechanism in SPAWN to integrate portions of MeSH. Basically, this has entailed making MeSH a parallel thesaurus of terms, and encoding a 'MeSH perspective' on terms in our thesaurus that have counterparts in MeSH. Furthermore, we are adding every term from MeSH categories A, B, and C that is not already in our thesaurus and classifying it according to MeSH structure.

Finally, we have expanded the natural-language processing potential of the thesaurus by 'linking' entries to our Project Lexicon. The Lexicon contains information about a term's synonyms and—through the use of a morphological analysis program developed during the Project—its morphological variants. Such information helps in mapping natural-language expressions to candidate thesaurus concepts. For example, the natural-language lexical items *abdominal, abdominal-region, abdomen, abdomens* and *belly* can all be mapped to the thesaurus entry *abdomen.* All lexical entries have pointers to their corresponding concept-terms in the thesaurus. However, not all concepts in the thesaurus are represented in the lexicon, so future work may involve expanding the Lexicon to include missing concepts.[18]

The sections that follow describe our work in constructing the basic classification hierarchy and developing complex clusters of concepts, as appropriate for the expression of *clinical findings.*

2.3. Constructing the Thesaurus

2.3.1. Activities Involved in Constructing the Thesaurus

The construction of the MedSORT-II Thesaurus was a long and, at times, tedious task that involved a great deal of concentration and reflection to maintain consistency among the semantic categories of the Bx hierarchy. This could not be done by a simple single-pass approach to the classification of terms. It involved a constant refinement of the hierarchy; the addition and deletion of both categories and terms; and the exhausting task of proofing the

[18] Some aspects of the Project lexicon are discussed in greater detail in Part III.

terms' categorizations for simple mistakes and subtle inconsistencies. We found ourselves continuously wrestling with semantic senses of individual terms and the intensional properties of their designated contrast sets. Should *systolic,* for example, be a *relative-index* or a *relative-measure* or a *pattern-quality?* Are all terms similar to *systolic* in the same category?

Since we endeavored to have only canonical forms of concepts in our thesaurus, relying on a morphological package to identify form-variants, we did not want to have multiple-form entries for the same concept. To assure this, we had to face the task of eliminating superfluous morphological variants (such as standard plural forms of already classified terms), and of removing or correcting of simple misspellings, and of choosing between alternate spellings. But it was not as simple as deleting all the morphological variants of a term. Very often, the adjectival form of a term had a different semantic role to play, and thus a different classification.[19]

Another major problem was the determining of what was considered to be an *atomic* concept, as we had to include among single words a great many multi-word phrases. Deciding what constituted an atomic concept was not straightforward. Would *acid-fast-bacterial-infection* be considered an atom? It might, afterall, be specific enough to be classified as a disease. Or would it be better broken down into parts like *acid-fast* and *bacterial-infection?* Some phrases were clearly atomic, such as *amyl-nitrite.* But even after it had been decided what constituted an atom,[20] the word ordering in the entry became a problem. In the INTERNIST-l/QMR findings, the word order of many atomic concepts was inverted, making it impossible for our parsers to recognise them as an atom (e.^., *strawberry-tongue* is TONGUE STRAWBERRY in INTERNIST-l/QMR findings). The decision was to preserve the order atoms would take in natural language, as that, and not the INTERNlST-l/QMR findings, was the target input for the parsers.

To maintain the consistency of the thesaurus entries, we first carefully classified them using medical dictionaries as references. The entries were then sorted by category, and the categories checked for consistency and correctness. Maintaining the conceptual consistency of categories such as *relative-temporal-index* and *ttmporal-relative-measure* was not an inconsiderable task, and even now, making precise the distinction between the two is difficult. Ver)" often, borderline cases were either doubly (or triply) classified, or moved back and forth between categories. With every major addition of terms, these processes were repeated, at least twice with the guidance of medical experts, who both checked our work and made suggestions about better distinctions to draw.[21] The end result is a thesaurus of about 5700 basic-concept entries referenced by a lexicon much larger than that.

[19] Generally, **all** adjectival forms **of** *body-parts* were classified as *body-regions,* as the adjective form referred **to the general region around the body part, not just the body part. For** example: *intestine vs. intestinal*

[20] **Our principal test for atoms was to ask whether the parts of a candidate term could stand alone meaningfully in the domain. For example, would *fast* (of *acid-fast)* play a role in structuring other concepts in biomedicine? We thought not.**

[21] **For example, our consultants suggested the addition of a *syndrome* category to capture those *pathological-statts* that were not quite diagnosed *diseases,* but were more specific than general states like *impaction.***

2.3.2. The Addition of the MARS Thesaurus

One of our goals was to expand our thesaurus to include terms from Neurology and Neuropathology. We chose the MARS Thesaurus as a source of candidate terms in this domain.[22] As our two thesauri are structurally incompatible, it was not possible to contemplate a simple merging of MARS terms into appropriate categories under the MedSORT-II Thesaurus.[23] Some of the relational/structural information in MARS, for example, could only be captured in our thesaurus by a decomposition of the MARS entry into one or more MedSORT-II Bx-level terms. Consider this point in somewhat greater detail.

The MARS Thesaurus contains at present 10047 entries. Each of these records at most seven fields of information, including the designation of such items as head-term, synonyms, variants, and general classification. There are also special fields called *narrower-than* and *broader-than* that play a special role in the cross-referencing of entries. Entries in the *broader-than* field are terms for which the head-term is *broader* semantically, i.e., capable of being used more generally or to refer to more things, *etc.* Entries in the *narrower-than* field are terms for which the head-term is *narrower* semantically. Since the MedSORT-II Thesaurus captures such distinctions in generality/specificity or primitive/composite by distinguishing Bx- from Cx-level terms, the only way we could hope to include the MARS *narrower-than* terms in the MedSORT-II Thesaurus would be to represent them as classes of Cx-level concepts. We did not have time, however, to conduct the proper analysis of MARS *narrower-than* entries required to develop appropriate Cx-level classifications, based upon the terms we have included as Bx-level concepts.[24]

As it was, we could only use the last field in the MARS thesaurus, the *pathology-class* field, to guide us in adding terms to the MedSORT-II thesaurus. But this field had not been carefully filled by those constructing the MARS thesaurus, and more than half of the thesaurus entries had *pathology-class* entries that were isomorphic to the head-term.[25] We chose to study only those terms for which there was a semantically useful *pathology-class* entry, an enumeration of which can be found in Figure 4.

After our analysis, we chose to work with the following pathology classes: *disease, anatomy, noun, adjective, procedure, manifestation, substance, physiology, animal,* and *etiology.* These were the largest classes with important entries that needed to be added to

[22] As noted earlier, the MARS Thesaurus is being developed at the Decision Systems Laboratory at the University of Pittsburgh School of Medicine, under the direction of Dr. John K. Vries, M.D. The thesaurus is only partially complete and not yet fully proofed at this time. We are deeply indebted to Dr. Vries for permitting us to study this resource. Naturally, he is not responsible for the accuracy of categorization or other use we have made of terms derived from our study of the MARS thesaurus.

[23] For more information about the MARS Thesaurus, see [Vries, *et al.* 1986].

[24] A preliminary study suggests that all of the MARS *narrower-than* terms can be recomposed from our **Bx-level** terminology.

[25] **The use of the head term to label the class of the head term is an artifact of the MARS Thesaurus development process: until properly classified, all terms are reduplicated to fill the *pathology-class* slot as defaults. The number of terms we found** in this state reflects the incomplete development of the MARS Thesaurus **at** the **time** we studied **it.**

The most frequent of the 4441 Pathology Classes used to categorize the 10047 MARS entries:

CLASS	# OF OCCURRENCES
disease	2185
anatomy	943
noun	778
adjective	471
procedure	353
manifestation	252
proper-noun	226
substance	219
physiology	75
animal	38
specialty	27
etiology	14
human	8
raw	8
organism	6
plant	5

Figure 4: Selected MARS Classes by Frequency

the MedSORT-II thesaurus that could give some guidance as to how the terms could be classified. This limited the number of entries we studied in the MARS thesaurus to 5220 *(Cf* Appendix B12.). Of these, 713 head-terms were common between the May 13 edition of the MedSORT-II thesaurus and the MARS thesaurus. Of these, everything that was given the pathology class *substance* in the MARS thesaurus and was not found in our thesaurus (a total of 167 terms) was classified and added to ours. Noting the amount of work it would take to classify all the entries we had chosen for study, and that a great deal of them were very complex concepts which could better be broken down, we then chose to take a careful approach to making further additions. At this point, we were left with a list of about 4300 entries from the MARS thesaurus to contend with.

The more careful approach we took was to examine these entries, limiting our interest to all single word head-terms and two and three word atomic head-terms, in the hopes that all the other entries would be constructed of these smaller units.[26] This was not a bad guess, for we found that there were only 372 single words (from the approximately 4300 single- and multi-word head-terms) that we had missed. In this way, a list of 2304 terms was generated that, when combined with terms from the MedSORT-II thesaurus, could be used to compose all 5220 entries we had chosen for study from the MARS lexicon *(Cf.* Appendix B13.). This list of 2304 terms was then to be classified according to our basic semantic hierarchy and added to the MedSORT-II thesaurus, enlarging it by about 45 percent.

Unfortunately, we found that the *pathology-class* entry from the MARS thesaurus could not give much guidance in our classification scheme, so we had to classify all the 2304 terms from scratch. Given time constraints, less than half the terms could be added to the MedSORT-II thesaurus, though we strove to make sure that the most important classes were added first *(Cf.* Appendix B14.). As it stands now, none of the terms in the pathology-classes *anatomy, adjective* or *noun,* or any of the 372 single words that had been neglected in our analysis of the head terms have been classified, and an attempt to classify these (about 1400) terms would take on the order of a man-week of work.

Work that remains to be done in the full integration of the MARS thesaurus includes studies of how the *broader-than* and *narrower-than* fields work. Specifically, we are interested in whether or not MARS head-terms that are broader than a large number of other head-terms (which would appear in its *broader-than* field) might not correspond to concepts at our Bx-level of analysis. Further, it may be that head-terms that are narrower than a large number of other head-terms (which would appear in its *narrower-than* field) would correspond to concepts at our Cx-level of analysis. Another method of generating Cx-level frames from the MARS thesaurus entries would involve studying co-occurence of terms from different Bx-hierarchy categories that were drawn from the MARS thesaurus. If regular patterns of co-occurrence were discovered in MARS head-terms—suggesting that the terms were genuinely composite—we could attempt to isolate new classes of Cx-level concepts in MARS. This method of analysis could not be undertaken as there was not time to fully categorize the MARS terminology. Clearly, the continued classification and addition of atomic concepts drawn from the entries in the MARS thesaurus is a job that must be completed if we are to

[26] This was based on the conjecture that a multiword MARS head-term could be constructed by concatenating entries in its *narrower-than* field.

say it has been fully integrated into the MedSORT-II thesaurus/knowledge base.

2.3.3. The Addition of the MedSORT-I Thesaurus

The addition of the MedSORT-I thesaurus required a special analysis of its structure so that it could be entered without gross re-classification of entries. This required both tracing the hierarchy of the MedSORT-I classification scheme and making sure along the way that terms referenced an existing node in the MedSORT-I hierarchy. The MedSORT-I hierarchy then had to be carefully restructured by the addition of higher level links so that it could be placed into the MedSORT-II hierarchy automatically, and so there would be no categories that dangled unconnected to the full Bx hierarchy. When this was done, the MedSORT-II hierarchy was greatly refined in the categories *macro-body-structure* and *macro-body-part,* among others. (See Appendix B6. for a full listing of the hierarchy.) Though the entries from the MedSORT-I thesaurus were thus very finely classified, the extra levels of refinement could not be used retroactively in the classification of the MedSORT-II terms as this would have required the reclassification of the major part of the MedSORT-II knowledge base.

2.4. Characterization of the Thesaurus

2.4.1. Explanation of Bx-Hierarchy Categories

In order to make clearer our sense of the denotation of categories used in the classification of the terms, we offer the brief descriptions of each of the principal classes given in the following Figures. These descriptions apply to each of the categories used in the MedSORT-II Bx hierarchy, but do not include the refined categories derived from MedSORT-I Thesaurus entries. Furthermore, we do not describe categories from the 'higher' levels of the Bx hierarchy, which are quite broad and general in their scope. It would be accurate to state that the categories we describe provide approximately the *basic-level distinctions* for clinical, diagnostic medicine. Virtually all the concepts in the Thesaurus have been classified under one or more of these specific categories. All examples are taken from the thesaurus.

As illustrated in Figure 5, the set of categories we identified to cover the broad class of human anatomy is tailored to the special needs of representing concepts in diagnostic medicine. As such, the critical distinctions fell in a realm completely different than the normal taxonomic principles of the medical field. At the highest level, all anatomic terms were considered as a *body-thing.* The highest level distinction then was between the concrete class of terms called *body-entity* and the more abstract class *body-region.* From here, distinctions among *body-part, body-structure* and *body-substance* were necessary. This division reflects the epistemological granularity we see in INTERNIST-l/QMR findings. See the figure for a precise explanation of each of the categories, and of the sub-classification of *body-region* into ***topological-body-region*** **and** ***relative-body-region.***

As shown in Figure 6, the set of categories we developed to classify concepts designating *pathological-factors* grew quite refined. This reflects the demands of diagnostic medicine, which require fine-grained specification of the pathologies afflicting the body. Furthermore, while the concepts used to define each category are clear enough, many terms could be

BODY-THING: anything generally related to the anatomy of the human body

BODY-EHTITY: a physical thing normally found within the body

BODY-STRUCTURE: any anatomical part that can be found in many places throughout the body, for which the term applies to all occurences, not a specific one (terms referring to specific structures are considered 'body-part's). The distinction between MACRO- and MICRO- has to do with whether the structure can be seen without microscopic magnification.

MACRO-BODY-STRUCTURE (e.g., joint, orifice)
MICRO-BODY-STRUCTURE (e.g., nucleus, rbc)

BODY-PART: any specific anatomical part of the body with a specific name, usually found only in one place. The MACRO-/MICRO- distinction is as above.

MACRO-BODY-PART (e.g., kidney, arm, head)
MICRO-BODY-PART (e.g., chorionic-villi, canaliculus)

BODY-SUBSTANCE: any substance not forming a particular part of the body, but which can be found in the body under normal conditions. (e.g., feces, breath, body-fluids)

BODY-FLUID (e.g., blood, serum, urine)

BODY-CHEMICAL: a chemical substance found in the body under normal conditions. The distinction between GENERIC- and SPECIFIC- has to do with whether or not the term can apply to a class of chemicals, or to a specific chemical substance.

GENERIC-BODY-CHEMICAL (e.g., hormone, enzyme)
SPECIFIC-BODY-CHEMICAL (e.g., acth, estrogen)

BODY-REGION: a terra that refers to a region of the body, not a specific part or structure. Any adjectival form of a part or structure was considered to refer to the region surrounding the specific thing referred to by the noun form. The distinction between TOPOLOGICAL- and RELATIVE- has to do with whether the region is tied to a specific part of the body, or is one that can be taken relative to many different parts of the body.

TOPOLOGICAL-BODY-REGION: See 'BODY-REGION'. The distinction between MACRO- and MICRO- is the same as for 'BODY-STRUCTURE'.

MACRO-BODY-REGION (e.g., abdomen, chest)
MICRO-BODY-REGION (e.g., extranuclear)

RELATIVE-BODY-REGION (e.g., acral, basal, dorsal)

Figure 5: Categories Included Under *Body-Things*

classified in several of the categories. For instance, there was no precise distinction between *syndrome* and *disease,* so some terms, such as *hypothyroidism* were categorized as both. Each of the categories is designed to identify a *contrast-set* of concepts, in which each term plays a similar role but counts as distinct from any other term. In general, the *pathological-factor* class is the single largest subsection of the hierarchy.

As illustrated in Figure 7, a number of important distinctions could be made among medical procedures. The most important was that between the *diagnostic-procedure* and *therapeutic-procedure* categories. As our focus was on diagnostic medicine, the *diagnostic-procedure* node wets most highly developed. We found that each of the different types of procedures listed did have distinct contexts and ranges of specificity. For instance, findings containing a *lab-assay-procedure* tended to specify pathological factors at the microscopic level—particularly concepts under *micro-organism* or specific *histologic-pathological-structure.*

Figure 8 presents a collection of categories under *physical-thing.* These nodes of the hierarchy did not require a great deal of refinement, as, for instance, in diagnostic medicine, *place* terms were generally found only in subordinate circumstantial information about the patient. The important distinctions under *substance* were between particular substances with specific medical purposes *(drug-substance* and *lab-procedure-substance),* and other general substances. Of the general substances, it became important to know which were generally found in the diet *(food-substance),* and those that were not *(non-food-substance).*

As illustrated in Figure 9, only a few simple distinctions were required under the organism node, corresponding to different infectious agents under the *micro-organism-,* and a limited collection of agents under *macro-organism,* that might be responsible for a bite or scratch.

Figure 10 presents the portion of the hierarchy devoted to an important abstract class, which we label *measure-theoretic-thing.* Concepts categorized here occur in evaluations, qualifications, and expressions of results of all kinds. A principal regularity among these specifiers and modifiers involves the the two axes *temporal/atemporal* and *absolute/relative.* A particularly important category under *quality* is the *evaluated-attribute* category, which collects those terms that are most relevant when associated with a particular measurement, such as *age* and *pressure.* As they are not fully interpretable without an associated measurement, including this class under *measure-theoretic-thing* is appropriate. While the taxonomy of the *quality* category may appear a bit odd, it reflects the grouping of terms we see among INTERNIST-l/QMR findings, biased necessarily toward diagnostic patterns.

The class of categories devoted to *experiences* and *circumstances,* given in Figure 11. is not very well refined, largely because of the secondary role played by information about patient subjective perception and circumstance in the formulation of clinically significant observations in diagnostic, internal medicine. Terms from these categories typically appear as modifiers of statements focused on pathological factors. For our purposes, these did not require detailed sub-categorization. In fact, the category *social-status* includes only one term *(homosexuality).*

A final, miscellaneous collection of categories under *abstract-thing* is given in Figure 12. These categories represent a set of semantically unimportant concepts—at least for the

PATHOLOGICAL-FACTOR: anything that is pathological to the body of the patient, anything 'wrong' with the body.

- PATHOLOGICAL-STRUCTURE: any structure found in the body that would not be found there under normal conditions. The distinction between HISTOLOGIC- and GROSS- has to do with whether the structure must be found by microscopic examination or not.

    - HISTOLOGIC-PATHOLOGICAL-STRUCTURE (e.g., megaloblast)
    - GROSS-PATHOLOGICAL-STRUCTURE (e.g., lesion, tumor)

- PATHOLOGICAL-STATE: any of a wide range of abnormal conditions that can afflict a patient, be they disease names or general conditions. (e.g., addiction, calcified, fibrosis)

    - SYNDROME: any pathological-state that is specific enough to deserve special denomination in the medical-domain, but is not as firmly established as disease (e.g., arthritis, hypothyroidism, adie-syndrome)
    - DISEASE: a specific pathological-state which may be considered a final diagnosis (e.g., aids, beriberi, diabetes-mellitus)

- PATHOLOGICAL-ACTION/PROCESS: any action that is abnormal or pathological to the normal functioning of the body (e.g., hemorrhage, vomiting)

- PATHOLOGICAL-SUBSTANCE: any substance generated by the body which, under normal conditions, should not be found there. The distinction between GENERIC- and SPECIFIC- has to do with whether the substance may be considered as a large class of substances and whether the term refers to a specific chemical or substance.

    - GENERIC-PATHOLOGICAL-SUBSTANCE (e.g., pus, exudate)
    - SPECIFIC-PATHOLOGICAL-SUBSTANCE (e.g., cryofibrinogen)

- PATHOLOGICAL-DETECTED-SIGN: a pathological factor that acts as an indicator for a particular disease or pathological-state (e.g., kerley-b-line, alexia, aciduria)

- PATHOLOGICAL-ETIOLOGICAL-FACTOR: a pathological-factor that may be considered the cause or etiology for a disease (e.g., neoplasm, toxin, infection)

- PATHOLOGICAL-FUNCTION: a normal bodily function which has ceased to perform normally (e.g., tachypnea, balding, polyuria)

Figure 6: Categories Included Under *Pathological-Factor*

MEDICAL-PROCEDURE: any action of any sort undertaken to either diagnose or treat a patient's condition.

DIAGNOSTIC-PROCEDURE: any procedure undertaken to try and diagnose what is wrong with the patient.

EXTRACTION-PROCEDURE: a procedure in which some sampling of either the patient's tissue or fluid is withdrawn for examination.

TISSUE-EXTRACTION (e.g., scraping)
FLUID-EXTRACTION (e.g., paracentesis)

LAB-ASSAY-PROCEDURE: a procedure requiring special equipment to examine the patient or his/her samples. A substance technique is performed on a fluid or involves the use of a special substance to obtain results.

SUBSTANCE-TECHNIQUE (e.g., serology)
TISSUE-TECHNIQUE (e.g., staining)

IMAGING: any procedure by which some part of the body is viewed, either by DIRECT- or INDIRECT- means.

INDIRECT-IMAGING (e.g., sonography, x-ray, nmr)
DIRECT-IMAGING (e.g., endoscopy)

MONITORING-PROCEDURE: a procedure whereby a bodily function is monitored, (e.g., ekg, eeg)

PHYSICAL-EXAM: any examination which involves physical contact between the doctor and patient and may be carried out in the confines of the doctor's office (e.g., pelvic-exam, percussion, inspection)

TEST-OF-DYNAMIC-FUNCTION: a test which focuses on the actions and reactions of a bodily function (e.g., cardiac-stress-test)

THERAPEUTIC-PROCEDURE: a medical procedure that attempts to treat a patient's ailment.

DRUG-ADMINISTRATION-PROCEDURE: any procedure that is used to give a drug to a patient (e.g., innoculation)

SURGICAL-PROCEDURE: a procedure involving surgery. (e.g., amputation)

PHYSICAL-THERAPEUTIC-PROCEDURE: a procedure by which a patient may attempt to regain lost abilities, (e.g., exercise-therapy)

Figure 7: Categories Included Under *Medical-Procedure*

PLACE: refers to a geographic location, be it an ABSOLUTE one, or one which is RELATIVE to some further specification.

RELATIVE-PLACE (e.g., home, valley, location)
ABSOLUTE-PLACE (e.g., mediterranean, ohio)

SUBSTANCE: any substance that is not generated by the body would fall into one of these categories.

DRUG-SUBSTANCE (e.g., dexamethasone, pentagastrin)

GENERAL-SUBSTANCE: a substance that is neither generated by the body, a drug, or used exclusively in some lab-procedure.

FOOD-SUBSTANCE: is something that may be found in a normal diet (e.g., meat)

VITAMIN (e.g., vitamin-c, riboflavin)

NON-FOOD-SUBSTANCE: is something not normally found in a healthy diet (e.g., asbestos, coal)

CHEMICAL-ELEMENT: any of the elements of the periodic table (e.g., carbon, bismuth, hydrogen)

LAB-PROCEDURE-SUBSTANCE: a substance that is used in a laboratory procedure such as a tissue-technique, or a skin test (e.g., coccidiodin, dexamethasone)

INSTRUMENT: a tool that is used for a specific purpose

MEDICAL-INSTRUMENT: any tool used in either a DIAGNOSTIC- or THERAPEUTIC-procedure.

DIAGNOSTIC-INSTRUMENT (e.g., balloon-catheter)
THERAPEUTIC-INSTRUMENT (e.g., prosthetic, pacemaker)

GENERAL-INSTRUMENT: any tool other than those used in medical procedures (e.g., cane)

Figure 8: Categories Included Under *Place*, *Substance*, and *Instrument*

```
ORGANISM: any living thing.

    MICRO-ORGANISM: any microscopic living thing.

        FUNGUS (e.g., phycomycetes)
        PROTOZOA (e.g., ameba, cercaria)
        BACTERIUM (e.g., streptococcus)
        VIRUS (e.g., herpesvirus, arbor-virus)

    MACRO-ORGANISM (e.g., dog)
```

Figure 9: Categories Included Under *Organism*

primary domain of biomedicine. The category *action/event* is potentially extremely important for the representation of *processes*, but INTERNIST-I/QMR findings do not capture cross-temporal conditions and contain virtually no verbs—and those that do appear are not critical to the sense of the finding.

The special class, *meta-language*, is used to categorize all names of categories. Terms that are so classified should (frequently) be co-classified under the category they designate or under the category that immediately dominates them. Consider the problem in giving the intuitively correct semantic interpretation to the phrase *heard unusual sounds in intestines.* Presumably *sound* is the category-name of such terms as *growl*, *ping*, *gurgle*, *etc.*, so what one hopes to be able to capture is the observation that *what* was heard was something that would be classified under *sound.* Do we reflect this when we write *sound isa sound*? Maybe, but we lose the distinction between *general* and *specific* sounds; and cannot recapture the additional information that comes from recognizing that *unusual* combines with *sound* to further restrict the possible specification of range of particular sounds. The problem here is that the correct interpretation derives from the composition of *unusual* and *sound*—*unusual-sound* is a sound. Thus, the general solution involves semantic typing, distinguishing generic from specific, and—most importantly—a rule sensitive to syntax:

> If a term that is classified as *meta-language* is modified by a term that is typed appropriately, combine the modifier and the meta-term to create a new term whose classification will be under the category designated by the meta-term.

This kind of problem, of course, presents itself in many contexts. In the phrase *no ratio of doctors to patients would have made a difference*, the term *ratio* is clearly *not* a ratio—a sequence of two numbers separated by ":", for example. But it is fair to say that *ratio* stands for a numeric-measure—*i.e.*, can be classified as *numeric-measure*, the category that immediately dominates it. Again, however, the real solution involves syntax, and the ability to distinguish generic terms and semantic types. The information we would like to capture is that whatever else the numeric-measure is, it has the form *<numeric-measure-of doctor-type>:<numeric-measure-of-patient-type>*. As long as we are merely exploring the classes of terms in INTERNIST-I/QMR-like knowledge bases, we can probably co-classify meta-terms under their named categories or the categories that immediately dominate their named

QUALITY: any descriptive information about a given thing.

SENSE-QUALITY: a quality of a patient's body sensed by a doctor in a physical examination (e.g., dry, hot)

PAIN-SENSE-QUALITY: a quality of a pain that the patient reports (e.g., aching, lighting, dull)

PHYSICAL-QUALITY: any information about the physical make-up of an object.

SHAPE-CONFIGURATION: the grossly observable physical shape of an object (e.g., round, bulging)
COLOR (e.g., red)
TEXTURE/STATE: the finer, sometimes microscopic, texture of an object (e.g., rough, tarry)

EVALUATED-ATTRIBUTE: a measurable or quantifiable something, usually a function of the body (e.g., size, age, pressure)

PATTERN-QUALITY: an adjective indicating a pattern of some sort (e.g., radiating, bilateral)

SOUND: a name for a specific sound or a quality of a sound (e.g., amphoric, click)

MEASURE: any descriptor denoting amount.

RELATIVE-MEASURE: any term which denotes a non-numeric amount. May be either TEMPORAL- or ATEMPORAL-, depending on whether or not it quantifies time.

TEMPORAL-RELATIVE-MEASURE (e.g., abrupt, recent)
ATEMPORAL-RELATIVE-MEASURE (e.g., abnormal, hot)

NUMERIC-MEASURE: any exact numeric amount.

RANGE (e.g., 50:300)
RATIO (e.g., percentage, csf/plasma)
INDIVIDUAL-NUMBER (e.g., 1, 25.96)

UNIT: any standard by which things are measured, being either TEMPORAL- or ATEMPORAL-.

TEMPORAL-UNIT (e.g., hour, second)
ATEMPORAL-UNIT (e.g., ml, lb)

RELATIVE-INDEX: an indicator of point, either TEMPORAL- or ATEMPORAL-.

RELATIVE-TEMPORAL-INDEX: a specific point in time, though it may be one that recurs according to a specific pattern (e.g., autumn, onset, systolic)

RELATIVE-ATEMPORAL-INDEX: a category reserved for directional modifiers that **were** not reserved to anatomy **(e.g.,** peak, left, right, front)

Figure 10: Categories Included Under *Measure-Theoretic-Thing*

PATIENT-EXPERIENCE: any condition that the patient might report during an examination.

PATIENT-PHYSICAL-SENSATION: any physical symptom the patient would report. (e.g., pain, tingling)

PATIENT-MENTAL-STATE: any mental condition the patient might report or that could be inferred from what the patient reports, (e.g., confusion, hallucination)

BEHAVIOR: actions that a patient performs, being either PATHOLOGICAL- (abnormal) or NORMAL-.

PATHOLOGICAL-BEHAVIOR (e.g., crying-spells)

NORMAL-BEHAVIOR (e.g. talking)

PATIENT-CIRCUMSTANCE: any circumstance that affects the patient

BACKGROUND-CONTEXT: any information regarding the patient's personal history.

SOCIAL-STATUS (e.g., divorced, married, homosexual)
DEMOGRAPHICS (e.g., male, oriental, european)
OCCUPATIONAL-STATUS (e.g., dockworker)

CONDITION-CONTEXT: any immediate circumstance affecting a patient's condition (e.g., at-rest, recumbent)

PHYSIOLOGICAL-CIRCUMSTANCE: any normal (i.e., not pathological) condition of the body.

PHYSIOLOGICAL-STATE: a state that is normal (e.g., awake)

PHYSIOLOGICAL-ACTION/PROCESS: an action or function that is normal to the everyday function of the body, or which indicates the body is acting normally, (e.g., bleeding)

PHYSIOLOGICAL-EVENT: a normal event in the functioning of the body, (e.g., extravasation)

Figure 11: Categories Included Under *Experience* and *Circumstance*

GRAMMATICAL-MARKER: terms that have little semantic meaning but are important markers syntactically.

QUANTIFIER (e.g., some, no, non)
CONJUNCTION (e.g., and, or, and/or, in-addition-to)

PREPOSITION (e.g., for, in}

ACTION/EVENT: any event that may happen to the patient (e.g., fall, bite, touch)

META-LANGUAGE (e.g., bacterium, temporal-unit <all previous classes>)

Figure 12: Miscellaneous Categories Included Under *Abstract-Thing*

categories. In fact, in the MedSORT-II Thesaurus, all meta-terms are co-classified under their immediately-dominating classes.

2.4.2. Excursions Into Sub-Classification and Tangling

Currently, the vast majority of entries in the MedSORT-II thesaurus are classified at the low-level nodes in the hierarchy explained above. Only a small fraction are more finely classified, and those are classified according to the MedSORT-I hierarchy, as a supplement to our core thesaurus *(Cf.* Appendix B17. and BIS.). Some items could not be finely classified and were therefore placed at higher levels in the hierarchy (i.e., *irregularity* is a *pathological-structure,* rather than either a *gross-* or *histologic-pathological-structure,* as it may be either). This would place most items at nodes found five-to-seven levels into the hierarchy. This provides a great deal of information that is critical for understanding natural language input, and allows for correctly parsing semantically meaningful input. But there is a problem in that the level of classification is not so fine that meaningless input would always be eliminated. When choosing among the set of terms under *macro-body-part,* a parser would give *toe* the same status as *descending-thoracic-aorta*—clearly undesirable if one could be substituted for the other in the input and the same result were returned. For this reason a number of efforts were made to better define the semantic roles of the terms in the Thesaurus.

Foremost among these were efforts to subclassify the terms meaningfully. This could be done in a great many ways, and it was pragmatic concerns that dictated the approach we took. Since we had done considerable and fruitful work examining the INTERNIST-l/QMR findings having to do with diagnostic procedures, and specifically *lab-assay-* and *extraction-procedures* and *imagings,* we chose this domain in which to try to subclassify the terms.

Our method was one that generated primarily ad-hoc classes of terms that were pertinent to those sets of findings. What was done was to examine all findings in which, for instance, a *fluid-extraction* term was found and to break it down into its constituent terms (entries in the MedSORT-II Thesaurus), and then sorting these terms by their categories *(Cf.* Appendix B16.). All terms in such a finding were then placed in a subcategory of their original category formed by appending the words *fluid-extraction* to the front of their category name *(i.e.,* if the term was *ascitic-fluid,* its category changed from *body-fluid* to *fluid-extraction-body'fluid).* An important note is that the term also retained its original categorization, so no information was lost. This allowed some restrictions on the slot-fillers of frames pertinent to *fluid-extraction* terms.

It is hard to say how well such a system can work, for some terms appear in a wide range of findings, and placing them in many subclasses at once seems a bit odd and allows the knowledge base to grow in size with no real gain in information. But, as it had already been necessary to allow tangling (the multiple classification of a single term in different branches of the hierarchy), and as no information was lost, there were minimal adverse effects on the knowledge base. A better approach, given unlimited development time, would have been to carefully study each category for delimitable subsets that would create usable restrictions for frame slots.[27] But determining these subsets was doubly difficult: not all categories are easy to break down; and the categories created may not add any information to the frames.[28] This would be counterproductive, since a principle of constructing the Bx hierarchy was to create semantically meaningful categories that would be useful in parsing natural language. Our method was a sure way of generating useful constraints, with the minimal addition of superfluous information.

2.4.3. The Addition of Other Relational Information to Bx-Level Terms

The addition of subcategories was only one way to add constraints to the frames used to parse input. Another method we used was to add other relations that formed a much more complex network than the comparatively simple Bx-z*5a-hierarchy. For all anatomy terms, we added information about any of three pertinent relations: *joined-by, part-of,* and *located-in.* And, to better restrict slot-fillers on lab procedure frames, we added a *location* relation for many specific laboratory procedures, including *imagings* and *extraction-procedures,* among others.

A term was said to participate in a *part-of* relation with another term if it could designate a functional part of the designatum of the other term. For instance, *colon* was identified as part-of *digestive-system.* The inverse relation of *part-of is has-part,* so we would write *"digestive-system has-part colon\'* A term was said to participate in a *located-in* relation with another term if it designated something physically located in the designatum of the other term. Note that *part-of* and *located-in* are different—a thing located in another need not play any part in the functional role of the other. For instance, the heart is located in the pericardium, but serves a completely different function. As the inverse relation of *located-in* is *contains,* the pericardium *contains* the heart. When possible, the most specific area was designated for the *located-in* relation, so that there may be several levels of things located in other things.

The *joins* relation, in distinction to *part-of* and *located-in,* takes three arguments, and

[27] Note that we were in fact not able to add restrictions to frame-slots during the time of the Project.

[28] More specifically, a slot that was formerly filled by category X may now be filled by any of X.I, X.2, X.3, all the new subcategories constructed from the terms in category X. If X.I, X.2, and X.3 contain all the terms X did, no information is gained.

was limited in application to joints. A joint was said to join two bones if it was the joint between then. For instance, the acromioclavicular joint was said to join the acromion and the clavicle. The inverse of the *joins* relation is *joined-by.* Either bone that is the object of the *joins* relation is said to be joined by the given joint, so the acromion is *joined-by* the acromioclavicular joint. The last of the relations we added was the *location* relation for laboratory procedures, which supplied the default *body-site* at which the procedure is usually performed. For instance, a bronchoscopy is performed on the bronchi, so has a *location* bronchi. These relations were not the only ones to be found in the MedSORT-II Thesaurus, as we included several more that were created for the MedSORT-I Thesaurus, specifically: *assoc-cell, assoc-discipline, assoc-proccess, bilateral, bones, connected-to, differentiation-product,* and *differentiation-source (Cf* Appendix B15.).

The general purpose of these additional relations was to supply critical information so that well-formed but meaningless input *(e.g., bronchoscopy performed on left foot revealed asbestosis)* could be eliminated and meaningful input could be recognized as such. These relations define alternative networks in the thesaurus/knowledge base that have potential implications for hierarchical organization of concept information and inheritance. Exploring the logical properties of combinations of links from this network is the next topic of this section.

2.5. The Logic of Relations in the Thesaurus

2.5.1. Excursions into Inheritance with Relations Other Than /5a

The relation central to the Bx hierarchy of the MedSORT-II Thesaurus is the *isa* relation, and its primary function is to allow for inheritance of information from high-level nodes to low-level nodes. The most important logical property that is involved in the inheritance is *transitivity,* whereby if *x isa y* and *y isa* z, then *x isa* z, and therefore can inherit the properties pertinent to z, as well as those pertinent to *y.* The inverse relation of *isa* is *inv-Lsa,* and would be used for tracing up the hierarchy to find more global categories. Clearly, it is also a transitive relation.

The *isa* relation, however, does not allow for the representation of other types of relations among terms. In the realm of anatomy, *located-in, part-of* and *joins* (and their inverses) all capture a different type of information than the simple *isa* classification. With these relations, however, assume transitivity. As the *joins* relation takes two completely different arguments (a joint and two or more bones), and these arguments must be given in a strict order whereby the joint is specified first, it can not be transitive. The *located-in* relation. however, does exhibit transitivity. If *x located-in* y, and *y located-in* z, then clearly *x locattd-in z* also. As such, information is inheritable along *located-in* links of the thesaurus.

More interesting, however, is the *part-of* relation. When taken without our definition, it may seem to be transitive, and in some particular cases it may be. But as we choose to define *part-of* by the functional role of the object of the relation (y, in *x part-of y),* it becomes clear that it is not a simple transitive relation. For instance, it would seem odd to say that the knuckle of the toe plays a role in the extremities, though tracing *part-of* and assuming transitivity, one would be forced to say that. It is not so much that this is completely incorrect to say such things, but that it is awkward to do so, and would require some explanation if the inference is drawn over large numbers of *part-of* links.

Where the study of these relations gets most complicated is when *part-of* and *located-in* links are considered in a path that combines the two. For instance, if *x located-in y* and *y part-of z,* then one could say that *x located-in z,* as one of the pragmatic conditions of something being *part-of* another is that it be physically contained by that other thing (it is assumed that a whole is the sum of its parts). Now the converse is not true. If *x part-of y* and *y located-in z,* it does not follow that *x part-of z.* There is no condition, pragmatic or otherwise, that says anything located in something else also be a part of the thing that contains it. Those are examples of binary combinations, and the complexity grows as more combinations are added. But it can be generalized that if *x located-in ṭ/, x located-in* anything found along paths of any number of either *located-in* or *part-of* links so long as the paths connect to *y.* This can be said as it is assumed that if *x part-of* y, then *x located-in* y. As can be seen, a great deal of information can be gained by inheritance, and knowing which relations permit inheritance in which combinations can be of critical importance to determining what information is available in a given semantic net.[29]

Another problem facing semantic networks is that of generic terms and specific terms, and an important distinction between what might be called *generic-generics* and *specific generics*—a generic term used in a generic sense and a generic term referring to a particular instantiation of that generic. For instance, blood as a generic can be thought of as a *body-fluid* that can be found throughout the body, but a particular part of blood, such as the myeoloblast, is better considered a *micro-body-part* and can only be found in bone marrow. At a higher level, even a gross sample of blood can only be considered to be have been found in one place in the body—the site from which it was taken—though one might wish to represent generic blood as flowing through all body parts. Thus the context of a generic term may play a very important role in how it can be classified in a semantic network. It has at least two distinct roles to play in language, and it is hard to encode this information in the classification of the term. This is a problem that we did not attempt to solve directly in the building of the MedSORT-II Thesaurus.

2.6. Suggestions **for** Futher Study

2.6.1. Analysis for Creating Additional Cx Frames

One of the byproducts of the building of the MedSORT-II Thesaurus was a number of intermediate concept clusters that could participate in even higher (Mx-level) clusters of concepts. The study of these clusters led to the creation of Cx frames in a number of areas pertaining to laboratory procedure findings, as reported in the following sections. But there remain many areas that have not been evaluated. It remains to be seen what clusters might be identified by studying the most complex of the MARS Thesaurus entries, and from a wide range of areas that were not within the realm of any of our sources of information.

[29] Detailed studies of the logic of semantic networks employing inheritance can be found in [Touretzky 1985], [Touretzky, *tt al.* 1986], [Thomason, *tt al* 1986], and [Horty, *tt al* 1986].

As an example of an area that could fruitfully be developed, a Cx frame, or perhaps several Cx frames, might be developed by studying how concepts cluster around terms in the Bx level category *Action/Event.* This would require the building in of temporal relations, but would add a great deal of information processing ability to the system.

2.6.2. Addition of New Structured Data

The whole of the MedSORT-II Thesaurus could benefit from further additions of data, particularly to refine some categories that we had no pressing need to subcategorize in our work. In addition, we were limited by time and other resources in pursuing the conceptual structure of the MARS Thesaurus. Only a relatively small portion of the terminology found in the MARS Thesaurus could be added to the MedSORT-II knowledge base. Since our use of MARS terminology depended on a semantic analysis compatible with that we developed for the core MedSORT-II Thesaurus, we faced the prospect of a potentially exhausting, parallel classification task. The addition of terms to the knowledge base by non-experts is an extremely lengthy and tedious process, and as the terms come from more and more specialized fields not found in standard medical dictionaries, it becomes all but impossible for non-experts to make the proper and necessary classifications. As this kind of development takes place, it becomes more necessary to have access to medical experts to check the non-experts' work.

2.7. Building the Mx and Cx Frames

With the Bx hierarchy in place, we proceeded to the task of developing representations of the *combinations* of concepts that play a special role in clinical medicine. As we noted previously, we used the analysis of findings-structure developed in the Carnegie Mellon University work on the UMLS Project as our model for the most complex combinations of concepts we would consider. The construction of findings-level frames guided us, as well, in the identification of intermediate, Cx-level concepts.

2.7.1. Hierarchical Organization of INTERNIST-I/QMR Findings

We chose to develop first detailed representations for laboratory findings; thus we focused on developing a frame for each node in the *lab-source-observation-mx* branch of the Mx hierarchy. Two main questions guided our efforts:

- What are the constituents of medical expressions that correspond to each node in the *lab-source-observation-mx* branch of the Mx hierarchy? That is, what are the *implicit* and *explicit* components of well-formed expressions of substance technique findings, imaging procedure observations, *etc.*?
- What are the semantic restrictions on these constituents? That is, which semantic categories from the Bx hierarchy correspond to each constituent?

The first question enabled us to identify the *slots* of each frame, while the second helped us to determine the *fillers* of each slot.

In order to address these issues and, thereby, create the frames for laboratory observations, we analyzed medical expressions that correspond to each leaf node of the Mx hierarchy—in particular, expressions of observations based on *assay procedures* (substance and tissue techniques), *extraction procedures* (fluid and tissue extractions), *imaging procedures* (direct and indirect imaging), *monitoring procedures, tests of dynamic function,* and other laboratory tests. The INTERNIST-l/QMR database once again provided us with a corpus of expressions for analysis, but these findings had to first be organized according to our Mx hierarchy. At the highest hierarchical levels, this work had already been done, since each INTERNIST-l/QMR finding was assigned to one of five major 'types': *sign, symptom, history, observation,* and *lab.*[30] Findings classified as *history* in INTERNIST-l/QMR were assigned to our *patient-history* category, *symptoms* to our *symptom* category, *signs* and *observations* to our *physician observation* category, and *lab* findings to the *lab-source-observation* branch of our Mx hierarchy. With these gross classifications in place, we turned to the more difficult task of sub-classifying findings according to the lower-level nodes in our hierarchy. We did this for two major classes of findings: *patient-history* and *lab-source-observation-mx.* Concerning the former, arriving at the sub-categories of *patient-history* and assigning particular findings to these categories was a cyclical process; the finding expressions themselves suggested sub-categories like *exposure, demographics, ingestion, family-history, etc.:*

- ***Exposure-Hx:*** **ARSENIC EXPOSURE HX**
- ***Ingestion-Hx:*** **DIET RARE BEEF INGESTION RECENT HX**
- ***Demographics-Hx:*** **RESIDENCE MISSISSIPPI OHIO VALLEY HX**
- ***Family-Hx:*** **OBESITY FAMILY HX**

Concerning the *laboratory-observation-technique-mx* findings, we relied upon the diagnostic methods expressed in the findings, since we used the affiliated methods as a semantic basis for structuring the *lab-source-observation-mx* branch of our hierarchy, in accordance with the advice we received from medical experts regarding the saliency of diagnostic method in expressions of laboratory observations. It was a straightforward matter to assign findings that contain only one diagnostic method to our sub-categories; we simply needed to know what semantic category a particular technique belonged to. For example, x-ray findings were assigned to the *lab-direct-imaging-mx* node. However, many INTERNIST-l/QMR findings cite more than one method. A frequent combination, for example, is an imaging procedure, a tissue extraction procedure, and a substance technique, as in BRONCHOSCOPY TRANSBRONCHIAL BIOPSY CANDIDA BY STAIN. Combinations such as these reflect a temporal ordering of medical procedures undertaken to derive a particular result. In our example, bronchoscopy was used to scan the lung for the particular site from which tissue would be extracted *(via* biopsy). Then, examination of the tissue sample by staining revealed Candida. For findings containing complex clusters of diagnostic procedures, we exploited this temporal ranking and classified them according to the procedure which most immediately yielded the result. Thus, our sample finding would be classified under *substance-technique-mx,* since

[30] *Lab* is further subdivided in INTERNIST-I/QMR into *labO, labl, Iab2,* and *Iab3,* according to a scale based on degree of invasiveness and expense, with *labS* being greatest.

staining most immediately revealed Candida. The other two techniques (bronchoscopy and biopsy) function as procedures that enable staining to be performed, and are therefore regarded as associated procedures by our system.

One feature of hierarchically-structured semantic networks used for natural-language processing such as the one we have been developing is that efficiency and accuracy of processing increases, up to a point, when finer sub-classifications are made. In terms of our Mx hierarchy, for example, nodes under *direct-imaging-mx* such as *x-ray-mx, arteriography-mx,* and *angiography-mx* could prove useful if exceptions to, or finer specifications of, values encoded in the *direct-imaging-mx* frame could be expressed in these subordinate frames. We could specify that the method of *x-ray-mx* is x-ray, that of *angiography-mx* is angiography, and of *arteriography-mx* is arteriography, and thus return an x-ray frame when an input string mentions x-ray as the detection method, rather than the more general *direct-imaging-mx* frame. Furthermore, a particular method may be combined with x-rays, say, that are never combined with any other direct imaging technique, and this restriction could be encoded in the *method* slot of the frame for *x-ray-mx.* We have not yet built frames at this finer, hierarchical level, but we have designed our Mx hierarchy so that further refinement along these lines is possible by sub-classifying *lab-source-observation-mx* findings according to their primary diagnostic method.

2.7.2. Analysis of Test Corpus of Findings

With each INTERNIST-l/QMR finding classified according to the *lab-source-observation* categories in our Mx hierarchy, we were ready to build frames for each category. In one respect, we approached this task 'bottom-up'—that is, we built the frames for the terminal nodes first, and then used these frames to build the frames for their 'parent' nodes. For example, we developed the frames for *substance-technique-mx* and *tissue-technique-mx,* and used these frames to create the more general frame for *lab-assay-mx.* More will be said below about how this was done.

In another sense, we took a 'top-down' approach to developing the frames, since we used the highest-level (and therefore most general) frame for *lab-source-observation-mx* that we had developed through consultation with medical experts and preliminary analysis of data as a basis for analyzing the findings. The *lab-source-observation-mx* frame can be represented as in Figure 13. Basically, this frame states that well-formed expressions of laboratory observations potentially include the following constituents, which are represented in the frame as slots:

- the primary *method* used to derive the observation;
- the medical observation or *result* itself;
- any other methods that were used, most likely as precursors of the primary method (which are included in the *method* slot);
- the specimen examined in particular laboratory procedures—in particular, lab assay and extraction procedures—which we labeled the *source;* and

```
(make-frame *lab-source-observation-mx
     (isa (value *observation-mx))
     (cases (value *method *source *location *result))
     (method (semantics *lab-observation-technique-cx))
     (source (semantics *body-thing-cx))
     (location (semantics *body-thing-cx))
     (result (semantics *pathological-factor-cx
                        *evaluated-attribute-cx
                        *quality-cx)))
```

Figure 13: The *Lab-Source-Observation-Mx* Frame

- the anatomical part or region that a laboratory technique was applied to—which we called the *location*.

To create the frames for the nine *lab-source-observation-mx* types, we performed a top-down, semantic constituent analysis of all INTERNIST-I/QMR findings that corresponded to each type. To facilitate this analysis, we organized the findings according to the Mx hierarchy that we developed, and paired each finding with its representation as a string of semantic categories from the Bx hierarchy, as shown in Appendix B7. Consider, for example, the endoscopy findings given in Figure 14, from the *monitoring-procedure-mx* section of the hierarchy.

Development of Mx frames proceeded as follows. First, each finding in each of the nine terminal sub-classes was labeled in terms of the top-level constituents of the *lab-source-observation-mx* findings: *method*, *location*, *focus*, *source*, and *result*. Note, for example, the constituent analysis of these *tissue-technique-mx* findings given in Figure 15. What this analysis shows us is that tissue technique findings primarily consist of a *method*, *source*, and *result*, although they might also include a secondary method, as in the case of LYMPH-NODE ASPIRATE YERSINIA FLUORESCENT-ANTIBODY STAIN POSITIVE. *Location* and *focus* are irrelevant, because tissue techniques are applied to specimens—most commonly a body fluid or tissue sample—rather than to anatomical structures, as are imaging techniques such as x-rays or arteriography. The frame for *tissue-technique-mx* would therefore be composed of these three slots.

As is evident in the examples, many constituents of findings correspond to phrases rather than to individual words. For example, the *result* of CSF SMEAR ACID-FAST BACTERIA, *acid-fast bacteria*, consists of two terms. The head term, *bacteria*, is modified by a type specifier, *acid-fast*. It was by further analyzing the sub-constituents of findings, like this, that we were able to derive the Cx frames—*i.e.*, the frames for semantic categories from the basic hierarchy. For example, the above analysis contributed to the development of the Cx frame for *micro-organism*. In particular, it showed us that well-formed descriptions of

- DUODENUM ENDOSCOPY ULCER CRATER <S>
  MACRO-BODY-PART DIRECT-IMAGING GROSS-PATHOLOGICAL-STRUCTURE
- STOMACH ENDOSCOPY DIFFUSE INFLAMMATION
  MACRO-BODY-PART DIRECT-IMAGING SHAPE-CONFIGURATION PATHOLOGICAL-STATE
- STOMACH ENDOSCOPY HIATAL HERNIA
  MACRO-BODY-PART DIRECT-IMAGING MACRO-BODY-REGION PATHOLOGICAL-STATE
- STOMACH ENDOSCOPY MASS
  MACRO-BODY-PART DIRECT-IMAGING EVALUATED-ATTRIBUTE
- STOMACH ENDOSCOPY MUCOSAL ATROPHY
  MACRO-BODY-PART DIRECT-IMAGING MACRO-BODY-REGION PATHOLOGICAL-STATE
- STOMACH ENDOSCOPY TELANGIECTASIA
  MACRO-BODY-PART DIRECT-IMAGING GROSS-PATHOLOGICAL-STRUCTURE

Figure 14: *Endoscopy-Mx* Findings

- <u>EYE <S></u> <u>PUNCTATE-KERATITIS</u> <u>BY ROSE-BENGAL STAINING</u>
  source *result* *method*
- <u>CERCARCIA SKIN-TEST</u> <u>POSITIVE</u>
  method *result*
- <u>SPUTUM</u> <u>SMEAR</u> <u>GRAM-POSITIVE DIPLOCOCCI PREDOMINANT</u>
  source *method* *result*
- <u>LYMPH-NODE</u> <u>ASPIRATE</u> <u>YERSINIA FLUORESCENT-ANTIBODY STAIN</u> <u>POSITIVE</u>
  source *secondary method* *primary method* *result*

Figure 15: Constituent Analysis of Tissue Technique Findings

micro-organisms potentially contain a qualifier, and that this qualifier can be of the type *physical quality.* We would represent this information in the *micro-organism-ex* frame as follows:

```
        (make-frame *micro-organism-cx
            (isa (value *organism-ex))
            (quality (semantics *physical-quality-cx)))
```

Further analysis might reveal other qualifiers, as was indeed the case with micro-organisms. For example, we found modifiers of quantity such as *numerous* as in *numerous rods*, and therefore added *atemporal-relative-measure-cx,* the semantic category of *numerous,* to the set of acceptable qualifiers in the *micro-organism-cx* frame:

```
        (make-frame *micro-organism-cx
             (isa (value *organism-cx))
             (quality (semantics *atemporal-relative-measure-cx
                                 *physical-quality-cx
                                 *quantifier-cx)))
```

All Cx frames were built up incrementally, in this fashion, by analysis of top-level constituents. The information in these constituents was encoded in twelve slots, which are described in Figure 16.

Often it was possible to generalize a value in a slot according to its specific values, as can be illustrated with the Cx frame for *substance technique,* the filler of the *method* slot in the *substance-technique-mx* frame. Substance techniques, such as skin tests, typically require an *agent*—a particular material that is instrumental in inducing a particular response. These agents may be *protozoa,* like *echinococcus* in *echinococcal immunoelectrophoresis-test, fungi* like *aspergillus* in *aspergillus precipitin-test,* or a *bacterium* like *chlamydia-group* in CHLAMYDIA-GROUP ANTIGEN COMPLEMENT-FIXATION-TEST. Initially, as we came across each category—*bacterium, protozoa,* and *fungus*—we listed that category in the *agent* slot of *substance-technique-ci.* But when the analysis of agents was complete, it became clear that we could consolidate tliese categories into the more general category, *micro-organism-cx.* and assign this category as a restriction on the *agent* slot. Such consolidation is important not only because it increases the readability of frames by making them more concise, but because it improves their performance in natural-language processing by enabling sub-classes other than those explicitly mentioned to be recognized as acceptable values of a given slot. For example, suppose that a type of *micro-organism* could function as the *agent* of *substance techniques* other than the three already mentioned. The parser would fail to recognize it as an *agent,* since its type is not encoded in the frame. Generalization to a higher-level category circumvents this problem.

Following semantic constituent analysis of findings, the second major stage of Mx frame development involved identifying the *restrictions* on values that can fill each slot. This was done in the same way as restrictions were identified for slots in Cx frames, as described above. Essentially, we listed all possible Cx categories that could fill each slot, and then consolidated these values into higher-level categories when doing so could improve natural-language processing performance.

location the anatomical site of a medical observation such as an *evaluated attribute, measure,* or *quality,* and of *body substances* (e.g., chemicals and fluids) and *body parts (e.g.,* particular cells).

agent that which induces the outcome of a *diagnostic method—e.g.,* the material used in *assay procedures (substance* and *tissue techniques).*

patient the anatomical entity which experiences a pathological event, such as the thing obstructed in the *obstructing-ex* frame.

quality a modifier on the central ('head') concept of a Cx frame—generally restricted to items of semantic category *quality, evaluated-attribute, measure,* and *relative index.*

context the situation in which an event *(lab source obsemation* or *medical procedure)* occurs, as in *rem during sleep.*

object in a strict linguistic sense, that which receives the action of a verb. For example, in the *containing-cx* frame, the object is the thing contained; in the *measure-cx* frame, that which is measured.

result the qualitative or quantitative outcome of a *measure, index,* or *evaluated-attribute.*

scale and quality used to specify a quantitative result. In *measure-cx* and *index-cx* frames, the *scale* is the unit in which a value is expressed, while the *quantity* is the number that precedes this unit.

source and goal used to express the scope of a physiological or pathological action or process. The *source* is the anatomical starting-point of this event; the *goal* is its terminal point, as in *pain radiating from chest to thigh. From* signals the *source (chest).* while *to* signals the *goal (thigh).*

instrument the medical device used in diagnostic procedures—*e.g., catheter* in *catheterization.*

Figure 16: Descriptions of Slots in Frames

The theme of consolidation carried over into the next phase of frame building. When all nine frames for terminal nodes in the *lab-source-observation-mx* branch were built, we created frames for their 'parent' frames, and for the 'parents' of these frames, etc. Doing so enabled us to revise the terminal Mx frames, since 'parent' frames were designed to encode default values that could be inherited by their 'children.' As an example, consider an early version of the Mx frames for the two types of *extraction procedures—tissue-extraction-mx* and *fluid-extraction-mx*—as given in Figures 17 and IS, respectively. Because both frames cite a *body-thing-cx* as a filler of their *focus* slots, we can represent this information in the 'parent" frame for *extraction-procedure-ex* and rely on inheritance for passing down the information to the *tissue-extraction-mx* and *fluid-extraction-mx* frames. Similarly, the *source* slot in the *extraction-procedure-mx* frame would be filled by *body-thing-cx,* and we would delete this slot from the *tissue-extraction-mx* frame. We would retain the *source* slot, with its value of *body-fluid-cx* in *the fluid-extraction-mx* frame, however, because this is the *only* type of body specimen used in fluid extractions, by definition. By doing so, we over-ride the default value of the 'parent' frame *{extraction-procedure-mx),* and ensure that only one type of anatomical entity is recognized during processing of natural language as an acceptable filler of this slot.

Identification of default values in the 'parent' frame, coupled with more precise specification of restrictions in the child frames, can also be seen in the *result* slot of our example. We see that both the *fluid-extraction-mx* and *tissue-extraction-mx* frames contain various *pathological-factor-ex* types in this slot, plus *measure-cx, quality-cx,* and *organism-ex,* so these restrictions can be encoded the the *extraction-procedure-mx* frame and deleted from the frames for *tissue-extraction-mx* and *fluid-extraction-mx.* But additional semantic categories are retained in the *result* slot of the terminal frames—such as *blood-count-ex, crystal-cx* and *micro-body-structure-cx* in the *fluid-extraction-mxfcame,* and *granuloma-cxm* the *tissue-extraction-mx* frame. The resulting, consolidated frames appear in Figure 19.

But *extraction-procedure-mx* findings are only one sub-type of a more general class of *lab-source-observation-mx* findings, the *lab-instrument-mx* findings, as can be seen in the Mx hierarchy, in Figure 2. Other members of this class are: *lab-monitoring-mx. lab-test-mx.* and *test-of-dynamic-function-mx.* The frames for each of these finding categories were used to create the 'parent* frame for *lab-instrument-mi.* given in Figure 20. which would, again. encode the default fillers of slots in its 'child' frames. With the default values specified at this level, the Mx frames for *lab-monitoring, lab-test, extraction-procedure,* and *test-of-dynamic-function* could be refined further. We see, for example, that it is no longer necessary to state that *body-thing-cx* satisfies the *focus* slot of *extraction-procedure-ex,* so this slot can be removed from the latter frame. Similarly, it would be redundant to include *measure-cx, pathology-ex*, and *quality-cx* in the *result* slot of *extraction-ex** since these values appear in the corresponding slot of the *lab-instrument-mx* frame. Consequently, we are left with the refined frame for *extraction-procedure-mx,* given in Figure 21. Appendix B8. contains all of the Mx and Cx frames, many of which have not yet been consolidated and refined.

```
(make-frame *tissue-extraction-mx
     (isa (value *extraction-procedure-mx))
     (cases (value *source *method  *focus *result))
     (source (semantics *body-thing-cx))
     (focus (semantic *body-thing-cx))
     (result (semantics *pathological-factor-cx
                        *quality-cx
                        *micro-organism-cx
                        *measure-cx
                        *micro-body-structure-cx
                        *infiltration-cx
                        *disease-cx
                        *organism-cx
                        •specific-body-chemical-cx
                        *action/event-cx
                        *body-substance-cx
                        *granuloma-cx
                        *physiological-action/process-cx
                        *deposit-cx
                        *formation-cx
                        *complex-cx
                        *degeneration-cx
                        *plugging~cx
                        *metajnorphosis-cx
                        *loss-cx
                        *filling-cx)))
```

Figure 17: An Early Example of a *Tissue-Extraction-Mx* Frame

```
(make-frame *fluid-extraction-mx
        (isa (value *extraction-procedure-mx))
        (cases (value *source *focus *method *result))
        (source (semantics *body-fluid-cx))
        (focus (semantics *body-thing-cx))
        (method (semantics *fluid-extraction-cx))
        (result (semantics *specific-body-chemical-cx
                           *measure-cx
                           *relative-index-cx
                           *quality-cx
                           *blood-count-cx
                           *suppression-cx
                           *pathological-structure-cx
                           •histologic-pathological-structure-cx
                           *pathological-subst2Lnce-cx
                           *micro-body-structure-cx
                           *action/event-cx
                           *response-cx
                           *pathological-state-cx)))
```

Figure IS: An Early Example of a *Fluid-Extraction-Mx* Frame

```
(make-frame •extraction-mx
     (isa (value •lab-instrument-mx))
     (cases (value •method •source •focus •result))
     (method (semantics •extraction-procedure-cx))
     (source (semantics •body-thing-cx))
     (focus (semantics •body-thing-cx))
     (result (semantics •micro-organism-cx
                        •pathological-factor-ex
                        •quality-cx)))

(make-frame •tissue-extraction-mx
     (isa (value •extraction-procedure-mx))
     (cases (value •method *result))
     (result (semantics *pathological-factor-cx
                        *measure-cx
                        *micro-body-structure-cx
                        *infiltration-cx
                        *disease-cx
                        *organism-cx
                        •specific-body-chemical-cx
                        *action/event-cx
                        *body-substance-cx
                        *granuloma-cx
                        *physiological-action/process-cx
                        *deposit-cx
                        •formation-cx
                        •complex-cx
                        •degeneration-cx
                        •plugging-cx
                        •metamorphosis-ex
                        •loss-cx
                        •filling-cx)))

(make-frame •fluid-extraction-mx
     (isa (value •extraction-procedure-mx))
     (cases (value *source •focus •method •result))
     (source (semantics •body-fluid-cx))
     (method (semantics +fluid-extraction-cx))
     (result (semantics •specific-body-chemical-cx
                        •relative-index-cx
                        •measure-cx
                        •blood-count-cx
                        •suppression-cx
                        •pathological-structure-cx
                        •histologic-pathological-structure-cx
                        •pathological-substance-cx
                        •micro-body-structure-cx
                        •action/event-cx
                        •response-cx
                        •pathological-state-cx)))
```

Figure 19: The Consolidated Set of *Extraction-Mx* Frames

```
(make-frame *lab-instrument-mx
     (isa (value *lab-source-observation-mx))
     (cases (value *method *source *focus *location *result))
     (method (semantics *lab-observation-technique-cx))
     (source (semantics *body-thing-cx))
     (focus (semantics *body-thing-cx))
     (location (semantics *body-thing-cx))
     (result (semantics *measure-cx
                        *pathological-factor-cx
                        *quality-cx)))
```

Figure 20: The Consolidated *Lab-Instrument-Mx* Frame

```
(make-frame *extraction-mx
     (isa (value *lab-instrument-mx))
     (cases (value *method *result))
     (method (semantics *extraction-procedure-cx))
     (result (semantics *organism-cx)))
```

Figure 21: The Fully Consolidated Frame for *Extraction-Procedure-Mx*

```
(make-frame *direct-imaging-mx
     (isa (value *imaging-mx))
     (method (semantics *direct-imaging-cx)))

(maLke-frajne *direct-imaging-cx
     (isa (value *imaging-cx))
     (header-for-mx (value *direct-imaging-mx)))
```

Figure 22: An Mx and Corresponding Cx Frame

2.7.3. Modifying the Frames for Natural-Language Processing

At this stage of development, the frames represented medical concepts (in particular, findings), but were not yet suitable for natural-language processing using RULEPAR. Three major types of information had to be added:

- information to accommodate the particular parsing style of RULEPAR
- information that would make information *implicit* in input phrases *explicit* in instantiated frames
- information that could enhance the system's potential to interpret user input accurately, and to recognize semantically unacceptable and ill-formed expressions

Each of these modifications will be discussed in turn.

2.7.4. Adapting the Frames to RULEPAR

RULEPAR works on the principle that in order for a frame to be instantiated, a salient 'token' (word or phrase) called a *header* must be present in the input. Since one of our primary aims was to map a user's expression of a finding to one or more Mx frames, we needed to encode information into the frames that would enable RULEPAR to 'decide' which Mx frames to instantiate. The semantically salient concept in Mx frames for laboratory findings is the *diagnostic method* used. Thus, in the Cx frames for methods that are named in the *method* slot of Mx frames—in particular, *substance-technique-cx, tissue-technique-cx. direct-imaging-cx. indirect-imaging-cx, fluid-extraction-cx, tissue-extraction-cx, test-of-dynamic-function-cx,* and *monitoring-procedure-cx*—we added a slot called *header-for-mx* which would trigger the corresponding Mx frame. An example, for *direct-imaging* frames, is given in Figure 22. The *header-for-mx* slot is thus similar to the *cx-header* slot in that it triggers a particular frame for instantiation.

Another feature of RULEPAR is that it relies on the presence of lexical 'markers' in frames to interpret the relationships between items in an expression. For example, the preposition *on* could signal that the noun it precedes is the object of a *diagnostic method*—the entity that the method acts directly upon—its *source* (in *extraction-procedure* findings, for example), or

```
(make-frame *substance-technique-mx
     (isa (value *lab-assay-mx))
     (cases (value *method *source *result))
     (method (semantics *substance-technique-cx))
     (source (semantics *body-substance-cx)
             (marker in))
     (result (semantics *measure-cx
                        •relative-index-cx
                        *pathological-factor-cx
                        *generic-body-chemical-cx
                        *micro-organism-cx)
             (marker for)))
```

Figure 23: The Modified *Substance-Technique* Frame, with Markers

location (in *imaging* findings, for example), as in the finding, *legionella direct fluorescent-antibody on sputum came out positive.* In order to enable the parser to recognize the role of *sputum,* we exploit the potential of lexical markers like *on* to signal the semantic role of words that they precede. We do this by encoding these markers in the slots for the semantic constituents that they signal. The modified frame for *substance-technique-mx,* in Figure 23, reveals how this is done. Note also how markers have been added to the *result* field, enhancing the parser's 'ability' to identify these constituents in expressions of *substance-technique* findings like the preceding *legionella direct fluorescent-antibody on sputum came out positive* and in *ameba-gel-diffusion was positive.*

2.7.5. Handling Implicit Information

One feature of natural language is conciseness; much more information is implied by an expression than appears in its surface form. Upon reading the phrase *light-chains in urine have increased* in a medical chart, a physician would automatically be able to identify the medical procedure involved, even though it is not stated. Physicians know that the only way to detect and measure light-chains in urine is by immunoelectrophoresis. The problem, from the computational linguist's point of view, is how to enable a system to do the same—especially systems like ours that are designed to map natural-language expressions of findings to canonical medical expressions such as those comprising the INTERNIST-l/QMR database. INTERNIST-l/QMR expressions almost always specify the *diagnostic method* associated with the finding, as in the INTERNIST-l/QMR finding that corresponds to the above example: **IMMUNOELECTROPHORESIS URINE LIGHT-CHAIN <S> ONLY INCREASED.**

Our approach to this problem has been to make implicit information explicit in Cx- and Bx-level frames. For example, we have encoded information about the *location* and *source* associated with diagnostic methods in the Bx frames for these methods, as shown in

```
(make-frame *arteriography
     (isa-cx (value *direct-imaging-cx))
     (location (semantics *artery)))

(make-frame *bronchogram
     (isa-cx (value *indirect-imaging-cx))
     (location (semantics *lung)))

(make-frame *ekg
     (isa-cx (value *monitoring-procedure-cx))
     (location (semantics *heart)))
```

Figure 24: Examples of Bx Frames Containing *Location* Slots

- EEG REM AT ONSET OF SLEEP
- SCHILLING-TEST WITH INTRINSIC FACTOR B12 ABSORPTION DECREASED
- STRAIGHT-LEG-RAISING-TEST POSITIVE
- BRUCELLA SKIN-TEST POSITIVE
- AORTOGRAPHY SUPRARENAL MASS <S>

Figure 25: Example INTERNIST-I/QMR *Method* Findings with Implied *Sources* and *Locations*

Figure 24. This information enables RULEPAR to instantiate the *location* and/or *source* slot of Mx frames, when such information is implicit in the method, as in expressions such as the INTERNIST-I/QMR findings given in Figure 25. We have only implemented this approach to implicit information in a small portion of our frames, but future work could be continued along these lines.

2.7.6. Refining Semantic Restrictions

Another feature of natural language that we wanted our system to be able to accommodate is restrictions on semantic co-occurence. For example, a ***body-part*** can not modify a ***lab-procedure-substance***, a ***vitamin*** can not specify the type of a ***micro-organism***, etc.. We addressed this issue to a large extent by encoding restrictions on the values of slots, and by adding rules to RULEPAR such as one that enables it to detect mismatched ***body-sites***. However, general semantic categories like those comprising the Bx hierarchy are often inadequate as restrictions on the ***agent*** slot of the ***substance-technique-mx*** frame. As we have

- **ASPERGILLUS [fungus] SKIN-TEST POSITIVE**
- **CERCARIA [protozoa] SKIN-TEST POSITIVE**
- **ECHINOCOCCAL [protozoa] SKIN-TEST POSITIVE**
- **TRICHINELLA [protozoa] SKIN-TEST POSITIVE**

Figure 26: Example INTERNIST-I/QMR *Skin Test* Findings

already seen, only certain types of micro-organisms can act as agents of particular *substance techniques* such as *skin-test* For example, the findings given in Figure 26 show that only *aspergillus, cercaria, echninococcus,* and *trichinella* can modify *skin-test. Virus,* which is also a type of *micro-organism,* is unacceptable in this context. The parser should be able to recognize this and return an appropriate error message. We have just begun to address this issue through *sub-categorization.* That is, we could sub-categorize *aspergillus, cercaria, echinococcus,* and *trichinella* as *skin-test-lab-procedure-substances,* and encode this value, instead of *micro-organism-ex* in the *agent* slot of *substance-technique-cx.* In this way, we could ensure more accurate interpretation of input. Future work will involve refining the values in frames to reflect this more detailed analysis of co-occurence restrictions.

2.7.7. Testing the Frame-Based Grammar

As always, the best way to detect omissions and inaccuracies in a frame-based grammar such as the one we have been developing is to test it with sample input. We therefore formulated a set of expressions that the parser should be able to interpret and for which it should be able to return an appropriate Mx frame. Appendix B9. contains these phrases, arranged according to the Mx frame that they should instantiate.

3. Exploring Semantic Classes and Relations

3.1. Introduction

This section describes work done in support of (1) the inductive construction of hypothetical semantic classes of terms in INTERNIST-l/QMR findings and relations between classes, and (2) the inspection and assessment of classes specified Hop-down' by a hierarchy of concepts constructed in consultation with medical experts. Much of this work was fragmentary and exploratory in nature, and is reported here only as providing examples of procedures useful in investigating the structure of medical terminology.

3.2. The Need for Semantic Validation of Categories

The Project's conceptual framework, largely due to work done in the UMLS Project, of a hierarchically organized system of 'basic' (i.e., term-level) categories forms the basis for progressively more abstract clusterings of concepts, first of basic categories into 'conceptual-cluster' categories, then of both basic and conceptual-cluster categories into 'manifestation' categories.[31] Such a systematic rendering of inter-related medical concepts aids in making explicit a number of important semantic relations which may be evoked only implicitly in many formulations of medical findings; this explicitation is a pre-requisite of natural-language processing of medical information.

The INTERNIST-l/QMR findings have been carefully crafted by medical experts both to display a greater regularity of surface structure than free-form medical descriptions, and—most importantly—to evoke the most salient and critical semantic and pragmatic relations involved in medical diagnosis. It is clearly of interest to use these INTERNIST-l/QMR findings both to improve and to test the proposed categorical scheme. An inductive, 'bottom-up' study of the lexical and syntactic patterns among findings allows lexical instantiation of the scheme's basic categories and the generation of potential new additions to the proposed framework. A deductive, 'top-down' study of the findings reinterpreted in light of the scheme's categories allows testing and validation of the proposed scheme. Such testing would in the long run address such questions as the following:

- Can a grammar of the findings that is based on the categories prove superior *(i.e.,* more economical, more plausible) to one based on intuitive, inductively-based word groupings?
- Do the natural higher-level constituents of such a category-based grammar largely conform to the proposed organization of manifestation- and conceptual-level categories?
- Are the most important relations, holding between the portions of findings which are assigned various categories, akin to the main semantic relations proposed?

The inductive work described below in Section 3.3.1. has assisted us in exploring the vocabulary and syntax of the findings and in fitting the basic categories to the specific vo-

[31] ***Cf.*** **Appendix Bl.**

cabulary of INTERNIST-l/QMR findings. In a subsequent validation stage, described below in 3.3.3., these basic categories were applied to the findings, and a small experimental grammar based on this application was constructed for a subset of the findings.

3.3. Work Done

3.3.1. Studies of the Lexicon and Word-Based Grammar of INTERNIST-I/QMR Findings

There are currently 4,101 findings, containing 21,461 word tokens. Their lexicon comprises 3,088 word types, of which 94 (3%) are numbers (either integers like ^5or fractions or ratios like *1:80).* Findings range in length from 1 word to IS words, with a mean length of 4.96 words. They are highly telegraphic extended noun phrases or proto-sentences, comprising expert medical terminology, and of an idiosyncratic syntactic structure only remotely related to that of standard English. For example, adjectival and modifying phrases often follow rather than preceed the nouns they modify *(e.g.,* FECES BLACK TARRY).

Single Words As in any investigation of a linguistic corpus, it was necessary to first extract all the words occurring in the findings, and consider their general characteristics. (This lexicon, sorted alphabetically and by decreasing frequency order, is presented in Appendix Cl. and Appendix C2.)

Hyphenated words are rare; only 26 out of the 4,101 findings contain such terms as **PARA-AORTIC, END-DIASTOLIC, ATLANTO-OCCIPITAL and HTLV-III.**

In the majority of cases, words are in the singular form, immediately followed by a marker of optional plurality enclosed in angle brackets (<S>, <ES>, <IES>, <AE>), *e.g.,*

- **ABDOMEN FLANK <S> HEAVY BILATERAL**
- **BURSA <AE> THICKENING GENERALIZED**

However, there are some exceptions to this general rule. The findings contain some plural **forms, such as BODIES, HALLUCINATIONS, FOLDS, FEET. IDEAS, CAPILLARIES.**

Other words and phrases are also enclosed in angle brackets; these can signal alternate expressions for some of the preceding terms and phrases:

- **ADH <VASOPRESSIN> PLASMA GTR THAN 6 PG PER ML**
- **COLD HEMOLVSIN TEST <DONATH LANSTEINER> POSITIVE**
- **BREAST <S> SKIN EDEMA <PEAU DORANGE>**

The relative frequencies of findings' parts of speech are distinctive. Some parts of speech occur with common frequencies. Prepositions are frequent: *by, to, with, of, in, after, at, without, into* are all in the set of 500 most occurrent words. A small number of conjunctions are also highly occurrent—*or, and* and *and/or* are among the 100 most frequent lexical items—though many other common coordinating and subordinating conjunctions *(e.g., but, because)* are absent. But many parts of speech are conspicuous by their absence. The most striking absences are (1) the lack of determiners *(a, the)* and demonstratives *(this, that)*, (2) the almost complete lack of conjugated verbs other than adjectival forms derived from verbs and passive verbal forms *(e.g., calcified, localized, relieved by rest, increased after secretin)* and (3) the lack of auxiliaries *(is, was, has, had, etc.)* and modals *(may, might, will, would, etc.)*.

The Length of Findings To search for preliminary groupings of terms, a useful procedure to follow is to examine a listing of findings ordered by increasing number of words, *i.e.*, by increasing length. Such a listing (presented in Appendix C3.) clearly shows which words can stand alone as complete findings in and of themselves (words which constitute 'sentences'), which modifying words can combine with them to form 2-word findings, and so forth. One examines the terms in findings of length *n*, groups them into rough, intuitively plausible categories, and then goes on to consider the terms in findings of length *n* + 1, asking if new terms fall into previously defined categories or if new classes of words appear.

The 60 words that can stand alone as a single one-word finding are overwhelmingly terms which, loosely speaking, describe psychological or physical pathological states *(e.g.*, ACALCULIA, AMNESIA, PARANOIA, INFERTILITY, PROTEINURIA, FEVER). If we examine those lengthier findings that contain an occurrence of these words, some highly tentative compositional hypotheses emerge.

1. Any pathological state can be terminated by the word HX to form a composite finding that the patient has a history of that disorder *(e.g.*, MEHORRHAGIA HX, METRORRHAGIA HX, PROTEINURIA HX). This is true of both the pathological state itself, and more detailed specifications of the disorder; thus we find all of the following findings centered around JAUNDICE: JAUNDICE, JAUNDICE FAMILY HX, JAUNDICE INTERMITTENT HX, JAUNDICE CHRONIC PERSISTENT HX, JAUNDICE OF PREGNANCY HX, JAUNDICE PRECIPITATED BY STRESS HX, JAUNDICE SECONDARY TO ORAL CONTRACEPTIVE HX.

2. Pathological states can be combined with both the corporeal location in which they are present and the procedure used to reveal their presence. Thus we find, in addition to TACHYCARDIA: EKG SINUS TACHYCARDIA, EKG SUPRAVETRICULAR TACHYCARDIA. EKG VENTRICULAR TACHYCARDIA.

3. Disorders may occur with descriptions of their temporal profiles, as in FEVER INTERMITTENT, FEVER PERIODIC EVERY SECOND DAY, AMNESIA POSTICTAL.

4. Disorders are related in various ways, among which we find

   - co-occurrence or non-co-occurrence: AZOTEMIA PROGRESSIVE WITHOUT PROTEINURIA;
   - the influence of one disorder by another: ABDOMEN PAIN EXACERBATION WITH COUGH.

5. Some disorders are such that they may be associated with a quantitative specification of the extent to which they are present: FEVER 41 DEGREE <S> C OR GTR, PROTEINURIA GTR THAN 3 GRAM <S> PER DAY.

The 247 findings composed of two words introduce some terms which are, intuitively, of different types. Here we first encounter

1. Parts and regions of the human body both large (SKIN, TONGUE, FACIES, PENIS, ABDOMEN, FEET) and small (RBC (red blood cell), WBC (white blood cell)). In two-word findings, these corporeal locations are associated with

   - disorders or abnormalities discovered in those locations: ABDOMEN TYMPANITES, **SKIN ACNE, PENIS GANGRENE, FACE PARESTHESIA, GINGIVA SWELLING. WBC ERYTHROPHAGOCYTOSIS;**
   - characteristics of shape or color which are not *prima facie* pathological, but nevertheless medically significant: FACIES LEONINE, OBESITY TRUNCAL.

2. Terms designating fluids or gases in or drawn from the human body (BREATH, SPUTUM, BLOOD, FECES, URINE), followed by usually abnormal characteristics of such substances (BREATH FRUITY, SPUTUM PURULENT, BLOOD INCOAGULABLE).

3. The name of examination techniques (ESOPHAGOSCOPY, SIGMOIDOSCOPY) and surgical operations (ADRENALECTOMY, MASTECTOMY). Examination techniques are followed, in two-word findings, by the name of some pathological condition discovered by means of the examination: ESOPHAGOSCOPY ULCERATION, SIGMOIDOSCOPY TELANGIECTASIA. In longer findings, examination techniques are also associated with more complex descriptions of detected disorders *(e.g.,* ESOPHAGOSCOPY DISTAL ESOPHAGUS LONGITUDINAL LACERATION <S>), but may appear in conjunction with other examination techniques when they are part of a multi-stage procedure of discovery (ESOPHAGOSCOPY BIOPSY GRANULOMA <S>, SIGMOIDOSCOPY ENTAMEBA HISTOLYTICA TROPHOZOITE <S> BY SCRAPING). Surgical procedures are often simply noted to have been performed in shorter findings (THROIDECTOMY HX). In longer findings, they sometimes play a role in describing the context in which some other abnormal condition occurs (MILK INTOLERANCE AFTER GASTRECTOMY HX).

   It is particularly clear for many examination and surgical technique terms that their morphological constituents indicate the region of the body involved in their performance (e.g. *colonoscopy, splenectomy).* Sometimes such an intra-lexically invoked body region is modified by another constituent in a finding. For example, in ESOPHAGOSCOPY PROXIMAL NARROWING *proximal* modifies the esophagus viewed by the doctor. Lexical decomposition might yield such information if it is not explicitly recorded in frames representing these terms.

4. Words standing for overall characteristics of individuals, which need not be abnormal (RACE, SEX, OCCUPATION). These terms are most often found in short findings, where they are followed by an appropriate specification of the value of that characteristic: **RACE ORIENTAL, SEX FEMALE, OCCUPATION MINER, OCCUPATION VETERINARIAN** OR ANIMAL HUSBANDRY. Some of these overall characteristics are very general, and represent wide classes of possible types of values *{e.g., behavior* in BEHAVIOR INTRUSIVE, BEHAVIOR SOCIAL WITHDRAWAL).

Inspection of the 578 findings of length 3 reveals, along with many new terms which fall into previously encountered categories, the following new types of words:

1. Those designating microscopic foreign bodies-unicellular organisms, bacteria, viruses and fungi *(e.g.,* CRYPTOCOCCUS, LEPTOSPIRA, BLASTOMYCES, BRUCELLA). These references are usually associated with a corporeal location or substance and a technique of detection particularly suited to such microscopic entities (BLOOD CULTURE **CRYTOCOCCUS, SPUTUM CULTURE FUSOBACTERIUM.**

2. Terms designating chemical substances which are generalized enough not to be clearly distinctive of human physiology *(e.g.,* CALCIUM, POTASSIUM, CAROTENOIDS, COPPER). These terms are associated with a bodily substance in which they have been found, and their quantity, or changes therein, may be specified (CALCIUM SERUM **DECREASED, POTASSIUM URINE INCREASED).**

Inspection of the 815 findings of length 4 suggest the following new classes of terms:

1. Terms designating drugs, typically associated with the word *administration* in a record of a patient's past medication: CAPTOPRIL ADMINISTRATION RECENT HX, HEPARIN ADMINISTRATION PROLONGED HX.

2. Phrases describing techniques of examining not the patient directly (like the previously encountered ESOPHAGOSCOPY), but rather substances withdrawn from the patient: HISTOPLASMA COMPLEMENT FIXATION POSITIVE, IMMUNOELECTROPHORESIS SERUM IGA DECREASED, ASPERGILLUS PRECIPITIN TEST POSITIVE.

3. The names of large, non-human organisms which have interacted with patients: BEE STING RECENT HX.

To determine whether the tentative classes suggested by shorter findings are reflected in the whole corpus of findings, it is useful to examine the more occurrent phrases found in the entire corpus.

Phrases and Multi-Word Lexical Items Category construction requires an easily accessible listing of common lexicographic contexts of individual words. This was provided by extracting all two- and three-word phrases from the findings. A list of the most occurrent of these elements of the findings' 'phrasal lexicon' is presented in Appendix C4.

The phrasal lexicon was put to use in two ways: (1) to assist inductive characterization of frequent word categories and constituents, and (2) to assist selection of word sequences which could be considered, in the context of the findings, to be single, indivisible lexical units.

(1) The phrasal lexicon suggests that the following roughly characterized word classes and groupings are highly occurrent.

1. Parts, regions, or substances of the human body, which may be designated by one or more continguous words: KIDNEY, HEART, LEG, LYMPH NODE, RESPIRATORY TRACT, BILE DUCT <S>, BONE MARROW, HAND SMALL ARTERY, HAND SOFT TISSUE, JOINT FLUID, PERICARDIAL FLUID, LEFT STERNAL BORDER. A frequent sequence of words designating a corporeal area on or in some body part is the sequence LEFT/RIGHT

UPPER/LOWER QUADRANT. It is notable that when such corporeal locations are designated by two or more contiguous words, (i) the semantic relation between the words is implicit, and (ii) the order of those words is quite free. Thus we find both LEG <S> SKIN (for *skin on/of the legs*) and SKIN NODE <S> (for *nodes in/made up of skin*), and both patterns like HAND SMALL ARTERY or SMALL INTESTINE, with an adjective preceding the noun, and MESENTERIC ARTERY SUPERIOR or INTESTINE <SMALL>, with an adjective following the noun it modifies. These corporeal location phrases occur at the beginning of the majority of the findings. When they do occur at the start of the finding (*e.g.*, SPINE INTERVERTEBRAL DISK CALCIFICATION) the first word in the phrase is typically a noun (like SPINE) rather than an adjective, but there are exceptions, such as ESOPHAGEAL LOWER SPHINCTER RESTING PRESSURE).

2. Parts, regions, and substances or fluids of the human body, associated with the name of some method or technique of studying them: KIDNEY BIOPSY, CHEST XRAY, HEART ANGIOGRAPHY, PELVIC EXAM, PULMONARY ARTERIOGRAPHY, BLOOD SMEAR. These methods may be composed of several words, *e.g.*, COMPUTERIZED TOMOGRAPHY <ENHANCED>, BIOPSY <OPEN>, BRUSH BIOPSY, and they may be concatenated: BRONCHOSCOPY TRANSBRONCHIAL BIOPSY, ENDOSCOPY BIOPSY, BIOPSY CULTURE (as in STOMACH ENDOSCOPY BIOPSY, LIVER BIOPSY CULTURE). When methods are associated with corporeal regions or entities, the word(s) designating the method often 'interrupt' (*i.e.*, are placed in the middle of) the words designating the body entity: ABDOMEN XRAY COLON, CHEST XRAY LUNG, CHEST PERCUSSION DIAPHRAGM.

3. Terms designating diseases, afflictions and dysfunctions are another salient class, usually associated with a corporeal entity or region, *e.g.*, OSTEOSCLEROSIS, MALIGNANT NEOPLASM, HEART MURMUR, ABDOMEN PAIN, IRREGULAR FILLING DEFECT, ABDOMINAL LYMPHATIC OBSTRUCTION. Here again, these 'pathological factor' terms can interrupt a body-location phrase: CHEST PAIN SUBSTERNAL.

4. Terms designating various bacteria and viruses, *e.g.*, STREPTOCOCCUS PYOGENES.

5. A highly occurrent class of phrases which refer to quantities of various kinds, and to changes therein: 40 DL, GTR THAN 20, NOT OVER 2:9 MG PER DL, 2 TIMES NORMAL, INCREASED MODERATE, INCREASED BY 110 PG PER ML.

6. A set of phrases which refer explicitly to the patient's history. These references always end with *HX*, preceded by a small set of modifying adjectives: ILLNESS HX, DISEASE FAMILY HX, SURGERY RECENT/REMOTE HX.

These patterns in the phrasal lexicon confirm that many of the patterns found in findings of length 4 or less are prevalent in the entire corpus.

(2) Another immediate use to which these lists of 2- and 3-word sequences were put was to determine which phrases were to be considered indivisible single units for the purposes of subsequent analyses (such phrases are henceforth referred to as "multi-words"). Clearly, if a word A is always followed by a word B in the findings, and the word B is always preceded by the word A, the phrase "A B" is a prime candidate for single-unit status, *i.e.*, for being soldered together into the single lexical item "A-B". If a word A always precedes a word B,

but B is preceded both by A and other words, the argument for creating a single "A-B" unit is weaker.

Multi-words were first generated using an a preliminary classification of terms, about two months before the end of the project. Using those 2-word sequences in the current 4,101 findings whose frequency was equal to or greater than 2, 3,469 candidate multi-words were generated. The generation algorithm required that, in a 2-word candidate, the probability of the second word given that the first occurred was 1.0; if that requirement was met, it then calculated the "backward sequential probability" that, given that the second word occurred, the first word would precede it. This "backwards probability" was assigned to the multi-word candidate as a tentative measure of promise, or "value". When there was a length-2 candidate with a 100% value of the form "A B", and another candidate, perhaps of a lesser value, of the form "B C", a length-3 candidate multi-word was generated of the form "A B C", and given the value of the "B C" candidate. This procedure was recursively repeated using previously created multi-words, thereby generating progressively longer multi-words of progressively weaker values.

These candidates were then compared to the 1,124 multi-words discerned by human judges and used in categorizing findings. At the time of the first comparison, of the 1,525 multi-word candidates generated by human judgment and/or by algorithm, only 67 (4%) were generated by both humans and the algorithm, 402 (26%) were generated only by the algorithm, and 1,105 (72%) were generated only by the humans. (The results of the comparison are listed in Appendix C5.) This very low level of agreement between the algorithm and human judges suggests that humans base their selection of multi-words on many more considerations than just relative transition probabilities. One consideration is clearly syntactic. For example, in the findings *material* always precedes *into,* but judges would not consider *material-into* as a proper multi-word because it does not constitute a syntactic unit. Secondly, even when an algorithmically proposed multiword does constitute a syntactic unit, humans will reject it if one of its constituents could vary greatly, even though it does not vary in the findings: for example, 25 MCG has a 100% algorithmic value, but numbers are too variable in general to make it acceptable. Finally, semantic and conceptual constraints are at work in human judgment: certain combinations of words belonging to distinct categories do not seem acceptable, *e.g.*. CHOLAXGIOGRAPHY COMMON DUCT has a 100% algorithmic value, but judges selected no multi-words which thus combined a part of the body and a technique used to examine it.

The candidate multi-words were considered by human judges, and a number of them were adopted. Thus, in the categorization scheme employed at the end of the project, the number of hypenated multi-words had grown to 1,579 (from 1,124 two months before) and So multi-words were generated or accepted by both humans and the algorithm (up from 67).

3-3.2. Lexically-Based Grammars for Subsets of Findings

Inductive work at the lexical level of the findings included experimentation with a lexically-based grammar (i.e., one which made no appeal to categories imposed top-down) of small sets of findings containing certain words. Among the sets considered were findings containing

(1) quantitative phrases, (2) the word *pain*, and (3) the word *blood.*

(1) The grammar of quantity phrases proved quite tractable, as such phrases comprise regular patterns composed of various kinds of units (*e.g.*, *liter, gram*), numerical constitutents of various forms representing numbers, ratios or ranges (*e.g.*, *25, 1:80, 2 to 5*), comparative adjectives (*greater, less*), prepositions and conjunctions. It was possible to distinguish such constituents in phrases such as 10 TO 25 MG PER DAY, GTR THAN 120 MCG PER DAY, NOT OVER 4 TIMES NORMAL. A small grammar for quantity phrases is provided in Appendix C6.

(2) Pain findings, such as POPLITEAL PAIN RELIEVED BY REST, CHEST PAIN SUBSTERNAL KNIFE LIKE OR TEARING, and ABDOMEN PAIN RADIATING TO RECTUM, were found to have constituents belonging to the following broad categories:

1. corporeal locations, which are specified either as particular parts of the body (SPINE) or as more imprecisely delimited regions (INGUINAL AREA, ABDOMEN, SUBSTERNAL, POPLITEAL (PAIN)), and are modified by relative spatial adjectives (ANTERIOR, LEFT LOWER QUADRANT, LATERAL),
2. pain quality specifiers (*e.g.*, STABBING, DULL, COLICKY, TEARING),
3. pain degree specifiers (*e.g.*, MODERATE, SEVERE),
4. time course modifiers (*e.g.*, PRESENT AT NIGHT ONLY, SEASONAL, ABRUPT ONSET, LASTING LESS THAN 10 MINUTES),
5. patient circumstance specifiers (*e.g.*, AT REST, FASTING),
6. influencing factor phrases, involving drugs or activities (*e.g.*, INDUCED BY EXERCISE, RELIEVED BY NITROGLYCERIN, EXACERBATION WITH ALCOHOL).

A small grammar for such pain findings is provided in Appendix C7. This exercise indicated higher order categories which would have to be considered in the future, *e.g.*,

- the abstract characterization of spatial patterns: EXTREMITY <IES> PAIN DIFFUSE, FACE PAIN LOCALIZED, ABDOMEN PAIN GENERALIZED, ABDOMEN PAIN GIRDLE DISTRIBUTION
- movement, change in location: CHEST PAIN SUBSTERNAL RADIATING TO BACK vs. MIGRATING TO BACK
- the complex interaction of time with other qualities and phenomena: CHEST PAIN PAROXSYMAL INCREASING IN DURATION AND/OR SEVERITY RECENT HX, EKG ST SEGMENT ELEVATION WITH RECIPROCAL DEPRESSION DURING SUBSTERNAL PAIN, ABDOMEN PAIN ONSET GTR THAN ONE HOUR POSTPRANDIAL

The exercise also gave us examples of ambiguity (*acute* is both a time-course and a severity indicator), and of the tug between functional and semantic classifications (*e.g.*, both ***exercise***, ***elevation*** and ***antacid*** can exacerbate or relieve pain, but seem different as activities, states and drugs, respectively).

(3) Finally, consideration of findings containing the word *blood* found three broad patterns:

1. <substance> *blood* <quantity/change in quantity> <influencing factor / patient circumstance>;
2. <examination technique> *blood culture* <pathogen + modifiers> <method of culture>;
3. *blood smear* <shape-specification> <pathogen + modifiers>.

It was encouraging to find that classes induced from the pain-findings could be reused so readily to analyse findings of a different type. No formal grammar was developed for these patterns.

3.3.3. Categorized Findings and Category-Based Grammars

During the descriptive analysis of the findings, Project members were constructing progressively larger and more elaborate versions of the category hierarchy, in consultation with medical experts. Many of the distinctions embodied in the hierarchy were conceptually based rather than resulting directly from patterns involved in the findings. For example, though the findings clearly indicate the pervasive importance of words and phrases designating "corporeal locations", the hierarchy's early distinction between *Body-Part, Body-Region* and *Body-Structure* was not suggested by the distinctiveness of the co-occurrence patterns of words like *liver, abdomen* and *annulus.*

It was therefore necessary at various stages of development of the category hierarchy, to classify the words in the findings according to the current scheme and consider what regularities where shown by the resulting "classified findings".

Categorizing the Findings In classifying the findings, one tags the words in each finding with each of their possible categories. Of the 5,100 terms categorized just prior to the addition of MedSORT-I terms, 354 (7%) were assigned to more than one of the existing 119 possible categories. The largest overlaps between categories are between:

- *Disease* and *Syndrome:* such terms as *myopathy,, syncope, coagulopathy, hepatic-fibrosis* and *mixed-connective-tissue-disease* are in both categories.
- *Disease* and *Pathological-Detected-Sign,* which share such terms as *mitral-valve-obstruction, ventricular-aneurysm, mitral-regurgitation.*
- *Pathological-State* and *Pathological-Action/Process,* which both comprise such terms as *myelinolysis. displacement, pallidoluysian-degeneration, reticulosis.*
- *Pathological-State* and *Patient-Mental-State,* since such terms as *amnesia* and *aphasia* are clearly members of both.
- *Pattern-Quality* and *Relative-Temporal-Index* share such temporal pattern descriptors as *daily, menstrual, diastolic, presystolic.*

Appendix C8. provides the details of ambiguous category assignments.

Since some terms have several categories, the expansion of all possible category assignments to a finding is combinatorial: that is, if a finding contains a term X with two categories, CXi and CX_2, and a term Y with two categories CY_a and CY_2, four alternative categorizations are produced, containing the four possible assignments CX1-CY1, CX1-CY2, CX2-CY1 and CX2-CY2. Such an expansion of all possible categorizations is uninformed by any notions of semantic constraints, and as a result some of the expansions may not be properly parseable.

During categorization, any sequence of words that has been categorized as a multi-word is detected and categorized accordingly. Categorization procedures search for the longest sequence of words that can constitute a recognizable multi-word.

Using the 4,101 findings, and using the above-mentioned categorization scheme *(i.e.,* prior to the addition of the MedSORT-I terms) that assigns one or more of 119 categories to 5,100 terms, the expansion of findings containing ambiguously classified terms produces 5,742 classified findings (a 28% increase). The classified findings are listed in Appendix C9.

The Category Phrasal Lexion Classified findings are made up of a sequence of words tagged by their categories. The search for regularities in sequences of words considered as sequences of categories is of course assisted by listings of 'category phrases' of various lengths. Listings of category phrases of lengths 2 and 3 are furnished in Appendix CIO.

When we consider categorized findings by order of increasing length we find the following regularities:

- findings of length 1 are overwhelmingly *Pathological-State;*
- findings of length 2 are for the most part concerned with the *Pathological-State* of various corporeal entities and attributes, but also provide demographic and historical information;
- findings of length 3 often elaborate on the preceding finding patterns, but introduce mention of bacteria and of the methods *(e.g., Extraction-Technique, Tissue-Technique. Indirect-Imaging)* used to determine findings.
- Longer findings introduce progressively more elaborated patient circumstances, influencing factors, evaluated attribute values, and conjunctions or disjunctions of the above-mentioned constituents.

In Figures 27, 28. and 29 we give a very small sample set of classified findings, ordered by increasing length, with decreasing frequency in each length.

These patterns are very similar to those originally found when inspecting the findings prior to their classification. Classification does not seem to reveal radically new configurations; but it does not contradict previously observed regularities either.

Manifestation Category Patterns To explore the patterns of co-occurrence of categories within findings, a multi-dimensional scaling analysis of (an early version of) categories was performed, based on a subset of the approximately 2,000 findings that are involved in

1 52 ***Pathological-State***

ACALCULIA

2 48 ***Macro-Body-Part Pathological-State***

GINGIVA SWELLING

CALF PAIN

2 26 ***Disease Background-Context***

ANGINA-PECTORIS HX

2 21 ***Pathological-State Background-Context***

HYPERTENSION HX

2 12 ***Macro-Body-Part Gross-Pathological-Structure***

SKIN PETECHIAE

STERNUM TUMEFACTION

2 11 ***Evaluated-Attribute Pathological-State***

RBC HYPOCHROMIC

SPEECH ECHOLALIA

2 11 ***Meta-Language Occupational-Status***

OCCUPATION FARM-WORKER

2 11 *Disease Background-Context*

WILSONS-DISEASE FAMILY-HX

Figure 27: Selected Classified Findings, Length 1 and 2

3 64 ***Body-Fluid Tissue-Technique Bacterium***

SPUTUM CULTURE ANAEROBIC-STREPTOCOCCUS

ASCITIC-FLUID CULTURE MYCOBACTERIUM-TUBERCULOSIS

3 33 ***Specific-Body-Chemical Body-Fluid Atemporal-Relative-Measure***

BILIRUBIN URINE PRESENT

CALCIUM SERUM DECREASED

3 29 ***Macro-Body-Part Grammatical-Marker Pathological-State***

EAR <S> BULLOUS-MYRINGITIS

3 25 ***Body-Fluid Tissue-Technique Fungus***

PLEURAL-FLUID CULTURE ASPERGILLUS

3 18 ***Macro-Body-Structure Grammatical-Marker Pathological-State***

SINUS <ES> TENDERNESS

3 14 ***Macro-Body-Part Gross-Pathological-Structure Grammatical-Marker***

SKIN **FIBROMA** <S>

3 12 *Macro-Body-Part Indirect-Imaging Disease*

HEART ANGIOCARDIOGRAPHY ATRIAL-SEPTAL-DEFECT

STOMACH BARIUM-MEAL PYLORIC-OBSTRUCTION

3 10 *Macro-Body-Pari Indirect-Imaging Pathological-State*

SKULL XRAY PLATYBASIA

Figure 28: Selected Classified Findings, Length 3

4 11 ***Macro-Body-Part Indirect-Imaging Macro-Body-Region Pathological-Action/Process***

SPLEEN RADIOISOTOPE-SCAN SPLENIC DISPLACEMENT

BRAIN COMPUTERIZED-TOMOGRAPHY CEREBELLAR CALCIFICATION

4 10 ***Macro-Body-Region Meta-Language Tissue-Technique Bacterium***

PERICARDIAL FLUID ANAEROBIC-CULTURE ACTINOMYCES

5 10 ***Specific-Body-Chemical Body-Fluid Atemporal-Relative-Measure Conjunction Individual-Number***

GLUCOSE PLASMA GTR THAN 300

HEMATOCRIT BLOOD GTR THAN 50

6 11 ***Macro-Body-Structure Body-Substance Tissue-Extraction Micro-Body-Structure Grammatical-Marker Atemporal-Relative-Measure***

BONE MARROW BIOPSY BASOPHIL <S> INCREASED

6 11 ***Macro-Body-Structure Macro-Body-Structure Tissue-Extraction Micro-Body-Structure Grammatical-Marker Atemporal-Relative-Measure***

BONE MARROW BIOPSY MEGAKARYOCYTE <S> ATYPICAL

7 43 ***Drug-Substance Action/Event Temporal-Relative-Measure*** *Preposition Temporal-Relative-Measure Pathological-State Background-Context*

ACETAMINOPHEN ADMINISTRATION PRIOR TO CURRENT ILLNESS HX

8 16 *Specific-Body-Chemical Body-Fluid Aiemporal-Relative-Measure Conjunction Individual-Number Aiempora/- Unit GrammaticaI-Marker AlemporaI- Unit*

CALCITONIN PLASMA GTR THAN 120 PG PER ML

Figure 29: Selected Classified Findings, Length 4, 5, 6, 7, and 8

pulmonary diagnosis. The measure of association used was based on the number of findings in which two categories occurred at least once, and represented the probability that one category would occur at least once in a finding given that the other had occurred at least once. (Details and results of the analysis are presented in Appendix Cll.) Two results were of interest. First, it seems that one can loosely organize categories along a dimension that goes from more easily observable body-external objects and substances *(e.g.,* general chemicals, waste products of the body, extracted body tissues, bacteria), through a dense central cluster of corporeal entities and their attributes and afflictions, to the difficultly observable microscopic body parts and psychological states. Secondly, the following clusters of categories were found to be the most tightly grouped:

- *Body-Chemical* and *Body-Fluid e.g.,* (ACID PHOSPHATASE) (SERUM) INCREASED
- *Action/Event* and *Relative-Temporal-Index, e.g.,* HEART MURMUR (SYSTOLIC) (EJECTION) LEFT STERNAL BORDER
- *Macro-Body-Part, Micro-Body-Structure* and *Tissue-Technique, e.g.,* (SKIN) (BIOPSY)
- *Pathological-State, Pathological-Action/Process* and *Macro-Body-Structures/Regions, E.G.,* (HEMORRHAGE) (GASTROINTESTINAL) ACUTE RECENT HX, LUNG BIOPSY (NECTROTIZING) (ARTERITIS)

A similar cluster analysis was performed about a month later with an expanded category hierarchy embodying finer distinctions and the full set of 4,101 findings. The previous arrangement of categories along an external/observable to internal/psychological dimension was no longer in evidence. The most tightly associated categories had become:

- *Bacterium* and *Tissue-Technique,* as in ASCITIC-FLUID (CULTURE) (MYCOBACTERIUM-TUBERCULOSIS), **(BRUCELLA) (SKIN-TEST) POSITIVE, URETHA DISCHARGE (SMEAR) GRAM NEGATIVE (DIPLOCOCIl).**
- *Body-Fluid* and *Specific-Body-Chemical:* this represents the previously noted "chemical in body-fluid" pattern.
- *Sound., Pathological-Detected-Sign* arid *Relative-Temporal-Indtx:* this cluster reflects the frequent association of temporally unfolding pathological symptoms like *heart-murmur* with some indication of a time-point or time-segment of their unfolding, as in **HEART MURMUR DIASTOLIC APICAL IMMEDIATELY AFTER S2.**

Though these cluster analyses did not propose any hitherto unsuspected groupings, they were valuable as a check of groupings suggested by the category phrase listings, since they reveal co-occurrence relations between non-contiguous constituents.

An Experimental Category-Based Grammar If regular sequences of categories can be found in the classified findings, they will inform the construction of higher-level conceptual clusters. To that end, a categorially-based grammar which will detect and label the most regular contiguous sequences of categories is called for; construction of such a grammar was started *(Cf.* Appendix C12.) In its present embryonic state, it recognizes the following complex constituents:

1. *body-site* phrases, made up of *body-parts*, *-regions*, *-structures* and their (numerous) modifiers;
2. *pathogen* phrases, formed by *bacteria*, *fungi*, *protozoa*, *viruses* and their modifiers.
3. *substance-in-fluid* phrases, made up of references to terms from *body-chemical* and *body-fluid*, and their modifiers (these phrases always have the meaning of "<*Body-Chemical*> is present in <*Body-Fluid*>")
4. *pathological-factor* phrases, made up of *pathological structures*, *states*, *processes*, and their varied modifiers
5. *knowledge-source* phrases, of the following types
   (a) *view*-phrases, denoting imaging techniques (*xray, bronchoscopy*)
   (b) *obtain*-phases, denoting techniques by which tissue and fluids are extracted from the body (*biopsy, by scraping, washing, thoracentesis*)
   (c) *preparation*-phrases, denoting techniques for preparing and examining extracted materials (*on gram-stain, by csf latex-agglutination*)

When these complex phrases, made up of contiguous elements, are unified, rearranged, and labeled, it is possible to better discern regularities in what is left unlabeled.[32] The preliminary grammar was just a first step leading towards examining regularities in

1. *evaluated-attribute* phrases, with their extremely varied dimensions (*density, excretion-rate, affect, lucency, acidity, adhesiveness, texture, waveform*) dictating (often in a complexly context-dependent way) the possible types and permissible ranges of values or positions of their dimensions;
2. *influencing-factor* phrases (*with neck flexion, after lactose ingestion, exacerbation by sensory stimulus*)
3. *condition-context* phrases (*at rest, postmenopausal, nocturnal*)
4. the mutual constraints and interdependencies of the above types of phrases and those which have already been successfully labeled.

For example, it was during the examination of categorized findings called for by the grammar that analysts remarked co-occurrence relations between various types of techniques. They seemed to reflect the following common-sense scenario involving temporal and material links between techniques:

- you may observe something (*e.g.*, by direct or indirect imaging) or
- you may extract it from the human body (perhaps while you are observing it, perhaps not), and if you do extract it

[32] Rearrangement of non-contiguous elements, like the two body-place phrases separated by *arteriography* in (CAROTID) ARTERIOGRAPHY (INTRACRANIAL ARTERIAL) SPASM was not done by the rudimentary grammar, in part because it was to be superseded by frame-based parsing procedures which could more reliably pluck out semantically related but syntactically separated constituents.

- you might transform it advantageously *(e.g.*, by culturing it)
- before testing it or examining it in some new fashion *(e.g.*, by fluorescence or staining)

Thus when the use of a technique like staining is mentioned in a finding, we can assume that an extraction technique was also employed and may or may not be mentioned explicitly, but the reverse does not hold.

Work on a grammar for the INTERMST-l/QMR findings halted about a month and a half before the end of the Project. It was felt that the syntactic idosyncracies of the findings warranted a shift to parsing phrases better reflecting natural language syntax. The locus of accommodation of co-occurrence constraints was shifted back from sentence structure to (i) sub-classifications (especially sub-classifications involving association with different techniques) in the lexicon and (ii) the addition to the lexicon of hierarchical relations derived from MedSORT-I, MARS and MeSH.

At present, it is apparent that the following semantic and pragmatic relations are of primary importance and require detailed representation, be it in the lexicon, in the hierarchy or in the grammar:

- the *contains* and *part-of* relations that organize anatomical sites, objects and substances;
- the relations imposed by various methods and techniques between (i) their 'input', i.e., the locations and entities they apply to, and (ii) their 'output', i.e., the objects, attributes and values which result from their application; these relations are sometimes well enough defined to allow 'chaining' of techniques with a consequential 'composition' of their constraints;
- the constraints imposed on the kinds of values that various attributes allow in specific contexts.

4. Managing Representations of Knowledge

4.1. Introduction

In order to deal intelligently with the enormous amount of data which has been collected and classified in the MedSORT-II project, it is important to be able to organize this information both in a strict hierarchical fashion, and in a situation-dependent cross-hierarchical fashion. In order to accomplish this, we have developed SPAWN, a system for representing real-world knowledge in frames. SPAWN allows us to define arbitrarily complex frames and networks of frames, hierarchically inherit values from those frames, update and query the knowledge base from many different viewpoints and constrain the creation of frames depending on information stored elsewhere in the knowledge base. The knowledge base is used in conjunction with RULEKIT/RULEPAR to parse English sentences describing medical findings—the knowledge base acts as a storehouse of semantic relations and restrictions on medical terminology which the parser uses, along with some simple grammatical rules, to posit a frame representation for the sentence.[33]

4.2. Objectives and Procedure

Our research goal in using SPAWN is to be able to represent information from many different sources in a unified and uniform fashion. As a bench-mark, we have developed a semantic hierarchy, and then classified terms taken from the INTERNIST-l/QMR, MedSORT-I, MARS and MeSH corpora according to this hierarchy. The resulting knowledge base was loaded into SPAWN. The terms from these corpora were loaded into their appropriate positions in the hierarchy. In addition, some new relations were added to the knowledge base, such as *contains* and *part-of*, in order to explore the possibilities of inheritance of relations. Finally, some work was done to explore the utility of *views* in SPAWN. A sample of the knowledge base which we have created is given in Appendix D3.

We felt that we needed to create very large knowledge bases, in order to prove both the generality of SPAWN and of our classification techniques, and to test significant chains of relations and interactions between relations. This will be discussed at length in Section 4.3.4. Furthermore, the *view* mechanism was tested by loading the MeSH hierarchy, which contained many terms which were common to the INTERNIST-l/QMR, MedSORT-I, MARS corpora. The *view* mechanism allows the knowledge base designer to specify that relations only hold between two nodes in a particular view, and not in general. Views are discussed in greater detail in Section 4.3.5.

4.3. **The** Facilities of SPAWN

This section will provide a general introduction to SPAWN, concentrating on its concepts and facilities, rather than on specific details of its implementation, design, functionality or

[33] A general introduction to frames and frame philosophy may be found in [Woods 1975] and [Brachman 1979]. A considerable variety of material relevant to knowledge representation may be found in [Brachman *k* Levesque 1985].

user-interface. These details will be covered in Appendix D2. and in Appendix D5. Briefly, SPAWN provides a tool for the creation and manipulation of real world knowledge, organized in a hierarchical fashion.

4.3.1. The Need for an Enhanced System

SPAWN allows a user to implement a knowledge base. In contrast with a traditional database, a knowledge base provides the ability to create a more complex network of knowledge. A knowledge base such as SPAWN allows the user to organize information in a highly interconnected fashion.

In contrast with previous knowledge representation systems, such as FRAMEKIT, SPAWN enables the user to relate *facets* hierarchically. It also allows *relations* and *views* to be treated hierarchically as well. SPAWN allows the user to store information at the level in the hierarchy where it is most relevant. With these features, the user can access and manipulate information more flexibly and efficiently.

4.3.2. Introduction to the Hierarchy

A *hierarchy* is a way of ordering things in a series, which contains levels, one above another. When information is placed in a hierarchy, it is usually arranged so that more general information is placed on the 'higher' levels of the hierarchy and more specific information is placed on the 'lower' levels. For example, in the biological sciences organisms are classified according to kingdom, phylum, class, order, family, genus, species and subspecies. In the animal hierarchy, then, the kingdom is considered the most general level and is therefore the top level. The most specific groupings of animals are subspecies and compose the lowest levels. However, any level of the tree may be treated as the top level if only that part of the tree is considered.

For some fields of study, scholars have created hierarchies of concepts relevant to the field. These hierarchies are commonly called *taxonomies*. Taxonomies of cognitive and social skills have been developed in the social sciences. Readers familiar with this type of classification may find it helpful to keep such examples in mind when dealing with the notion of inheritance.

4.3.3. Frames

All information entered into SPAWN's knowledge base must be included in the hierarchy. SPAWN allows the user to organize information hierarchically by creating a network of frames.

Frames provide a way of representing knowledge. A frame may stand for an idea, or represent a class of ideas, a class of objects, or an actual physical object, depending on the way in which the knowledge base is being used for a particular application. For convenience, we will refer to the various entities being represented as *concepts*. Frames can be used to store information about the properties of a concept, and to express relationships between that concept and various other concepts.

```
[make-frame  *body-part
    (isa
        (value  *body-entity
    (inv-isa
        (value  *macro-body-part  *micro-body-part))]
```

Figure 30: A Sample Frame

For example, a *body-part* is a *body-entity.* The two kinds of *body-part* are *macro-body-parts* and *micro-body-parts.* SPAWN represents this information using the frame given in Figure 30. In this frame *body-part* is the *subject.* The *relation* between *body-part* and *body-entity* is represented by the words *isa* followed by the value ^*body-entity* which is said to be the *object* of the relation. The inverse relation of the isa-relation, or *inv-isa,* signifies that both * *macro-body-part* and * *micro-body-part* are *body-parts.*

4.3.4. Types of **Relations**

SPAWN provides for different kinds of system and user-defined relations. The most commonly used are *unidirectional* and *bidirectional* relations. Unidirectional relations link one concept to another but point only in one direction. To relate the same concepts in the opposite direction one must use a relation that is defined as inverse to the first. For example, one can define a *contains* relation to assert that veins contain blood. The inverse of the *contains* relation, *is-contained-by* can be used to assert the opposite fact—that blood is contained by veins.

A *bidirectional* relation has an inverse relation that is identical to itself. For instance, *next-to* is bidirectional: the incisor tooth is *next-to* the canine tooth, which implies that the canine is *next-to* the incisor.

Another possible kind of relation is a *save* relation which may be specified of another relation and keeps track of every place in the hierarchy where that relation is used. This makes it possible for the user to retrieve a list of all nodes that are linked by the relation to which the *save* relation is attached.

In addition, relations can be defined as *inheritance* relations—*i.e.*, relations which hold over any number of links. The *isa* relation is the most commonly used inheritance relation. For instance, in the MedSORT-II hierarchy, a *generic-body-chemical* *isa* *body-chemical* which *isa* *body-substance,* and so on. By following *isa* links in this hierarchy, we can thus discover long-distance relations.

Figure 31 shows how SPAWN hierarchically represents its relations. When a user adds relations, they must be added to this hierarchy. This is a very flexible technique for defining hierarchies of relations. For instance, we have defined an *isa-ex* relations which links Bx frames with Cx frames. The *isa-cx* relation was added to the relational hierarchy as a logical son of the *isa* relation. The *isa-cx* relation links head finding concepts in the Bx hierarchy with their Cx frames. So, for instance, there are *isa-cx* links between the *tissue-technique* and the *i'issue-technique'*cx frames, and between the *substance-technique* and the *substance-technique-cx* frames. Consequently, through the hierarchical inheritance mechanism, when SPAWN is queried to find out what Cx frame a term like *smear* is linked to, it follows *isa* links to the *tissue-technique* frame, and then follows the *isa-cx* link to the *tissue-technique-cx* frame. The query, however, would only ask for the *isa-cx* links of *smear*, and SPAWN would know to follow any number of *isa* links, until it finds an *isa-cx* link. To reiterate, SPAWN knows to do this because of the hierarchical definition of relations.

4.3.5. Views

Views alter the way in which one views the knowledge base. A view provides a window that restricts the way in which the knowledge base designer can create, link, examine, and retrieve information. For example, an osteopath looking through a large medical knowledge base w^r^ould want to 'see' more and different information about bones, such as links to bone diseases, than would a general practitioner.

The default view in SPAWN is called the *common view.* The *current view* influences how information is entered into the knowledge base. Once information is entered, the *reference views* influence how it is found. New views may be defined hierarchically and the user may change between them at will. Consequently, information (frames) may be present in one view and absent in another, or simply linked differently in the two view^r^s. This gives the user the flexibility of having many different related knowledge bases with the overhead of only a single knowledge base, plus a few extra links.

4.3.6. Inheritance by Relation

Inheritance allows concepts to inherit values from the nodes to which they are related. These nodes may reside higher up or lower down in the hierarchy. When the user relates one concept to another through *isa* links, SPAWN implicitly knows to maintain inheritance in the opposite direction and does so by calling on procedures known as *demons.* Demons can be written for any other relations, at the user's discretion.

Note that there are complications when retrieving information about concepts which are subordinate to a relation that is bidirectional. In these cases where the bidirectional relation is the subject of the inquiry, the subordinate relations are also treated as bidirectional, even if they are represented within their own frames as unidirectional.

In some instances the user may wish to override inheritance. For example, one concept may possess many properties, all of which are inherited by its subordinate relation except one. In such cases one who is knowledgeable of LISP may write general demons that cause the default inheritance to behave differently.

```
[make-frame ^relation
    (isa
        (value ^concept))
    (invJsa
        (value *inheritance-relation *inverse includes *facet *attribute))]

[make-frame ^inheritance-relation
    (isa
        (value *relation))
    (invJsa
        (value *isa))]

[make-frame *isa
    (isa
        (value *inheritance-relation))
    (inverse
        (value *invJsa))]

[make-frame *inverse
    (isa
        (value *relation))
    (inverse
        (value *inverse))
    (if-added-by-slot
        (value inverseSvalue$if-added-by-slot_demon))
    (invJsa
        (value *inv_isa))]

[make-frame xinv_isa
    (isa
        (value ^inverse))]

[make-frame xincludes
    (isa
        (value ^relation))
    (valueSif-needed
        (value includes-valueSif-needed-by-slot_demon))]
```

Figure 31: SPAWN'S Relational Hierarchy

4.3.7. Non-Inheritance Relations

Typically, some of the relations used in a knowledge base simply relate a value or attribute to its parent frame, and so there is no need for, and no possibility of, inheritance. For instance, in our Mx hierarchy, relations such as *assoc-method, source,* and *rank* statically link frames together, without recourse to inheritance of information.

4.3.8. Viewed Inheritance

The properties that a concept inherits depends largely on the view in which the concept is viewed. For example, lupus w^r^ould be considered by a rheumatologist to be a rheumatic disease and consequently in the rheumatologist's view it would inherit those properties included in a *rheumatic-disease* frame. However, in the skin specialist's view lupus would be considered a skin disease and thus would inherit properties from a *skin-disease* frame.

4.3.9. Viewed Inheritance of Relations

Within different views the hierarchy of relations may be defined differently such that in each view a relation may inherit different properties. For example, in one view a relation may be bidirectional and thus have an inverse which is equal to itself while in another view the same relation may be unidirectional and have an inverse which differs from itself.

4.3.10. Facets

Facets name the kinds of properties that fill the slots in a frame. For example, in the body part frame given in Figure 30 the facet is the name *value.* The other facet which is typically used in the implementation of the MedSORT-II Thesaurus is *semantics.* Like everything else in SPAWN, facets may be defined hierarchically.

4.4. SPAWN Functions

SPAWN'S functions are discussed at length in Appendix D2. The description of SPAWN'S functions includes, when appropriate, a discussion of what each function does, those situations in which the user would want to invoke the function, how to use the function, and any relevant feedback which SPAWN provides.

Of special interest is the SPAWN-loop function. SPAWN-loop is designed to help users learn how to use SPAWN. SPAWN-loop acts as an interface between the user and SPAWN'S internal routines by providing a syntax which is often easier to use than the syntax that SPAWN requires. However, SPAWN-loop is limited in that it offers only some of SPAWN'S commands. SPAWN-loop is described at length in Appendix D2..

Part III

Natural-Language Processing Applications

1. Introduction

An important, secondary goal of the MedSORT-II Project has been to develop methods for utilizing thesaurus-based resources in the processing of actual natural-language expressions. This part of the Final Report presents a discussion of the natural-language processing facilities we developed to test applications of the Project Thesaurus and knowledge-management system. Section 2, *Morphological Analysis,* describes our approach to the problems of handling morphological variation and lexical standardization. We discuss, in particular, MORPH, the Project morphological analysis program. Section 3, *Parsing Natural Language,* reports on our principal demonstration of the utility of the Project thesaurus/knowledge base, the processing of natural-language findings expressions. We describe our modifications of RULEPAR—RULEPAR-II—and offer examples of the interaction of the parser, MORPH, and **SPAWN.**

1.1. Natural-Language Processing

A great deal of research has been conducted on the problems associated with the automatic processing of natural language. This Report will not attempt to review past work or propose particular approaches to natural-language processing (NLP). However, we take for granted that one of the critical lessons of several decades of NLP research has been that the general problem of language understanding—and language representation—cannot be solved by attention to one or two aspects of natural language, in isolation from other natural-language phenomena. We cannot hope to process language by attending only to syntax, or semantics, or lexical representations: or by elevating sentences to a higher status than discourses. Our challenge is to develop systems of representation that facilitate gracefully the integration of many linguistic phenomena—and many levels of linguistic interpretation. In our work, in particular in our development of the MedSORT-II Project thesaurus/knowledge base, we have been guided by this view.

1.2. Implications for Biomedical Informatics

We have also been advised, however, by the lessons that have accrued in work that has focused on *limited* and *specialized* domains of discourse, where the linguistic structures that encode the relevant information may be unusually 'regular' or predictable, compared to unrestricted natural-language discourse domains.[34] Clearly, reports of clinical findings, whether

[34] *Cf.* [Grishman & Kittredge 1986] and [Sager, *et ai* 1987] for discussions of the characteristics of NLP in restricted domains, including biomedicine.

in hospital charts or research articles, will exhibit a great deal of regularity, and may be susceptible to NLP techniques that are presently suitable for large-scale applications. We would suggest that our experiments—in the parsing of clinical, *laboratory-source* findings—demonstrate not only that our Thesaurus has achieved a useful, integrated design, but also that the automatic processing of technical, biomedical information, in natural-language form, is in fact currently feasible.

2. Morphological Analysis

2.1. Introduction

The parser of the MedSORT-II Project, RULEPAR-II, requires a lexicon to provide the syntactic and semantic information associated with lexical items in natural-language expressions. We have developed a Morphological Analysis Package (MORPH) to recognize morphological variants of entries in the Project Lexicon and to provide the parser with relevant syntactic and semantic information for those morphological variants. In this section we describe the objectives and rationale of developing MORPH, the details of how it performs morphological analysis, and the lexicon which has been built to serve MORPH and, as a consequence, **RULEPAR-II.**

2.2. Objectives

There are two principle objectives in developing a facility to handle morphological variation. The first is to reduce the amount of morphological redundancy in the lexicon since—in English and most languages—morphologically related words usually share the same features. We can do this by listing only canonical forms in the lexicon, and deriving the morphological variants of these canonical lexical entries through application of morphological rules. The second is to automatically generate syntactic and semantic information for a given lexical entry according to its syntactic category. The syntactic and semantic information is used by RULEPAR-II to do natural-language processing. The category information for a canonical entry is listed in its feature list in the lexicon; the category information for an inflected or derivational form of a canonical entry can be derived from its affix(es).

2.3. Procedure and Rationale

One task for the natural-language processing in the Project is to match a user's input with terms in the Thesaurus. The user's input is a string of words which may contain morphological variants of thesaurus terms. For example, a term classified as *macro-body-part* is usually a noun. In the thesaurus, we typically use the singular form of the noun to denote it. So, *e.g.*, *hand* is in the thesaurus while *hands* is not. In order to match the word *hands* in input with *hand* in the thesaurus, we need MORPH to map *hands* to *hand.*

To index the thesaurus accurately, we need to have a lexicon that includes all its canonical terms. When a recognizable word is encountered in input, it should be identified as a term in the Bx hierarchy and the corresponding frame should be instantiated. For example, if we receive the word *jaw,* we can note that it is in the lexicon; and since it bears the feature *val bx*—telling us that it is a thesaurus term—we can invoke its associated frame from SPAWN. But entries like *hand, jaw,* and *superior-artery* are often used in their plural form, as *hands, jaws* and *superior-arteries,* which are not thesaurus terms and are not associated with frames. When we come across words like these, we want to reduce them to their singular forms so that frames associated with their singular forms will be instantiated. Thus, when we have the word *hands,* we want MORPH to recognize it as the plural form of *hand.*

The following sections describe types of morphological variation that the input may contain and two principle ways of dealing with such variation.

2.3.1. Morphological Variation

The two types of affixes which produce morphological variants are are termed *inflectional* and *derivational.* Inflectional affixes indicate some syntactic feature of the word they are attached to *(e.g.,* person, number, tense, aspect); they never change the syntactic category of their stems. Derivational affixes, by contrast, sometimes change the syntactic category of the stem they attach to. For example, if the suffix -er is attached to a verb *(e.g.,run)* it changes the word to a noun *(e.g., runner).* On the other hand, if the suffix -er is attached to an adjective *(e.g., long,* the resulting word is still an adjective *(longer).*

Some inflectional markers are completely regular, and stripping off the productive affix produces the canonical or stem form of the word. For example, the productive third-person singular marker is -s; for the word *eats,* the suffix *-s* can be stripped off, and the resulting form is the infinitive form *eat.* In other cases, regular inflectional markers may be attached to stems that have been transformed from their canonical forms in some way. For instance, the word *empties* contains the affix *-s.* When the affix is stripped off, the resulting form is *emptie,* which does not match the canonical form *empty.*[35] Thus some more analysis must be done on *emptie* before it is identical to the canonical form *empty.* As another example, the word *beginning* contains the suffix *-ing.* When this is removed, the resulting form is *beginn* which contains a final geminate (or doubled) consonant. Before this form can be matched with the canonical form *begin,* the final consonant must be degeminated. As a final example, consider the word *longer.* This word contains the suffix -r, which when stripped off produces the stem *longe.* The final -e must be removed to produced the canonical form *long.*

Still other types of morphological variation show greater irregularity and do not lend themselves to the sort of morphological analysis suggested above. For instance, some affixes are not productive, occurring on only a few forms in the language, *e.g..* the plural suffix *-en* in *children.* Some involve a stem-internal change, so that the suffix cannot simply be stripped off. *e.g..* the plural *of foot* is *feet,* the plural of *man* is *men.* In some cases a plural suffix other than *-s* is used, in particular for words of foreign origin, *e.g., media* as the plural of *medium, phenomena* as the plural of *phenomenon, stimuli* as the plural of *stimulus,* and *matrices* as the plural of *matrix.*

2.3.2. Possible Approaches to Morphological Variation

There are two major strategies for treating the morphological variation found in the input to a natural-language processor. One involves, in essence, ignoring the relationships between morphologically related words and instead treating every word as a separate lexical entry. In this approach all words having the same stem will nonetheless have separate listings in the lexicon. An advantage of this is that the input can be directly matched to stored representations, with no computing necessary prior to the attempted match. The implicit claim

[35] Note: for verbs, the canonical form is identical to the infinitive.

is that all lexical entries have equal status, regardless of whether they are morphologically simple or complex, and regardless of whether they are free or bound morphemes.

There are several disadvantages to this approach. One stems from linguistic considerations: significant generalizations are missed by listing all morphological variants; as a result, there is a great deal of redundancy in the lexicon. There are two types of generalization we might want to capture: one concerning *stems* and the other concerning *suffixes.* A stem is related in both form and meaning to all of its morphological variants. If we take this generalization into account, we can reduce redundancy in the lexicon by listing a stem, with all of its syntactic and semantic features, only once and by noting elsewhere *(e.g.*, in a morphological analyzer) that the variants are related to this stem and thus share those same syntactic and semantic features. For example, the word *altering* is a morphological variant of the stem *alter.* From its form (<stem> + *-ing)* we know that it is the present participle of the verb *alter.* And we can infer that it shares the same features as its stem.

A suffix also has consistent syntactic and semantic properties regardless of the stem it is attached to. Of particular interest here is that a suffix carries information about the syntactic category of the stem and also of the inflected or derived word. Thus, the category of a morphological variant can be determined by examining its suffix. For example, the word *abruptly* is a morphological variant of the stem *abrupt.* The suffix *-ly* attaches to an adjective and creates a derived form, an adverb. Thus, a morphological analysis identifies the suffix, matches the stem to an entry in the lexicon, and determines the category of the derived word.

Another disadvantage of listing all variants in the lexicon is that the lexicon will be very large. This will potentially cause two problems in computational applications. The first is that larger lexicons occupy more space in memory. On most current work stations, memory is quite limited; and a lexicon the size of the MedSORT-II Lexicon would be difficult to load. If loaded, there might be insufficient space to run programs that would utilize it. Even if we choose to leave the lexicon on disk and maintain an index to the lexicon in memory, there will still be problems, as we describe in detail below. The second problem is that a huge lexicon, with entries having similar feature lists, poses a problem for lexicon maintenance. If the feature of one entry needs to be updated, all of its morphologically-related entries would have to be updated, too. This increases the probability of mistakes in lexicon editing and validating.

The other strategy for dealing with morphological variation is to record only canonical or stem forms in the lexicon, wherever possible,[36] and to have a morphological analyzer identify affixes, record the category of the word containing the affix (and other grammatical information), and strip off the affix to identify the stem—which then can be matched in the lexicon. One advantage of this approach is that it is more sophisticated linguistically; it maintains relationships between morphologically related words and reduces the amount of redundancy in the lexicon. Another advantage of having a morphological analyzer is that, since only canonical forms are listed in the lexicon, maintenance of the lexicon is easier and less vulnerable to error—the addition of new items requires only one new entry (the

[36] **This is typically possible for items that show regular 'productive' morphology.**

```
(function cat n rel-to cytology subval mesh)

(function cat v subcat intrans subval mesh)
```

Figure 32: Lexical Entries for *Function*

canonical form), with its syntactic and semantic features, rather than several (one for each variant). A disadvantage of this approach is that more computation is required, because direct matches will not always be posible. More precisely, in such cases an attempt to find a morphological variant in the lexicon will not lead to a direct match, so computation time will be required to generate forms that can be directly matched to stored entries.

2.3.3. The Treatment of Morphological Variation in the MedSORT-II System

The approach to morphological variation taken in the MedSORT-II Project is close to the second one outlined above. The MORPH package[37] includes a morphological analyzer which recognizes a (limited) set of suffixes and reconstructs the canonical form of words containing those suffixes. It matches words to entries in the lexicon, and, for all words (even morphologically simple words), it returns a list of syntactic and semantic features associated with that word. In the sections that follow, each component of the MORPH package is described. Later sections offer detailed descriptions of how MORPH operates, with examples of output.

Morphological Analyzer. The morphological analyzer receives input which has been put into a standardized form by an initial filter. MORPH, in conjunction with the lexicon, maps recognizable forms (both canonical forms and variants) to lexical entries. It returns complete lexical entries and also adds some grammatical feature values based on the syntactic categories of the entries and their morphological form. For example, as shown in Figure 32, the word *function* is listed in the lexicon both as a noun and a verb. MORPH matches the word *function* with both of entries and adds grammatical features according to their categories. So the output feature list[38] of the word *function* from MORPH would be as given in Figure 33.

When morphological variants are recognized by MORPH, some grammatical features are also added to the feature list of their corresponding stem form. For example, the word *indicate* is listed as a verb in the lexicon and has the feature list shown in Figure 34. After the word *indicated* is recognized as the past form and past participle of *indicate*, the features *form (past pastprt)* and *png (s1 2 sm sf sn p1 s3)* are added to the feature list of the word

[37] Two versions of MORPH are described below. One is the COMMON LISP version, which tailored to run on the Hewlett-Packard "Bobcats" of the Laboratory for Computational Linguistics. The other is the FRANZLISP version, which runs on a Vax 11/780. In principle, they are the same, and will be described below simply with reference to MORPH unless otherwise stated.

[38] For a complete description of the features included in the Project Lexicon, please refer to Appendix E2.

```
(sense function cat n rel-to cytology subval mesh png (sm sf sn))

(sense function cat v subcat intrans subval mesh form (pres inf)
        png (si 2 pi p3))
```

Figure 33: The MORPH-Analyzed Lexical Entries for *Function*

```
(indicate cat v subcat intrans trans val bx)
```

Figure 34: Lexical Entry for *Indicate*

indicate, giving the feature list in Figure 35.

Lexicon. The current lexicon contains all the terms used in the Thesaurus. It includes the category information of each entry, as well as syntactic and semantic information that cannot be generated by MORPH, such as *syn* and *rel-to.* It also contains pointers to other resources such as the MeSH lexicon. An example of a lexicon entry is given in Figure 36. It includes the entry itself and its syntactic category. It also may include features that relate the entry to another entry or entries.

Two interesting properties of this lexicon are the inclusion of bound morphemes *(e.g., a=} dys=)* and phrasal terms *(e.g., acute-leukemia, resistance-to-flexion, Wilson's-disease)* as separate lexical items.[39] The inclusion of bound morphemes (both prefixes and suffixes) increases the scope of the Lexicon without the cost of increased lexicon size. For example, a term like *hyperextension* does not need to be listed as a separate entry in the lexicon because we have both the canonical term *extension* and the bound-morpheme *hyper=* and can compose them to produce a representation of the unlisted term.[40]

The inclusion of phrasal items makes the matching of input more accurate and reduces ambiguity in parsing. For example, *Wilson's disease* is treated as a disease-name instead of

[39] For phrasal items, the hyphen has been introduced as a joiner between atomic concepts; the entire hyphenated item is considered to be the lexical item.

[40] The Lexicon is designed to facilitate the composition of bound- and free-morphemes as described in this example, but we have not implemented this ability in MORPH.

```
(sense indicate cat v subcat trans val bx suffix d form (past pastprt)
    png (si 2 sm sf sn s3))
```

Figure 35: The MORPH-Analyzed Lexical Entry for *Indicated*

```
(lymphatic cat adj subcat qualitative syn (lymph-vessel lymphatic-vesel)
     rel-to lymph val bx subval mesh)
```

Figure 36: The Lexical Entry for *Lymphatic*

```
(adherent-pericarditis cat n rel-to pulsus-paradosux)
```

Figure 37: The Feature List for *Adherent-Pericarditis*

a disease-name modified by a determiner. In this way, phrasal and sentential ambiguities are greatly reduced. Once a continuous string of terms is recognized as a phrasal term (by a filter[41]), it will continue to be treated as an atomic lexical item. For example, the phrasal term *adherent pericarditis,* listed in the lexicon as *adherent-pericarditis,* has the feature list shown in Figure 37.

Figure 38 provides general statistics on the composition of the Lexicon, including information about the number of bound morphemes and phrasal lexical items.

Index. To be effective, MORPH must be fast. A major concern is whether its accompanying lexicon should be accessed directly from the disk or loaded into memory. The srtii Lexicon currently has about 11,000 entries; and it will grow apace as more concepts are added to the thesaurus. Even on large workstations, there may be insufficient space to hold the entire lexicon; or to leave room for other application programs, the whole lexicon is loaded successfully into the memory, it may prevent MORPH and the parser from working. The solution to this problem is to leave the lexicon on the disk but reduce the time required to identify matches between the input and the lexical entries. Since access time is fixed, we need to limit the number of times (per input word) that the disk must be accessed in order to identify a match.

This is achieved in MORPH by building an index to the lexicon.[42] In the index, every entry is paired with its position on the disk. A hashtable is used to store the index so that no searching is involved in matching an entry in the lexicon. Once the index is loaded, all analysis of a word from input can be done without accessing the disk till the word is recognized as a direct match or a possible variant of an entry in the lexicon. The disk is accessed at most one time for each possible variant, and the entry and its features are loaded into memory. A portion of the index is given in Figure 39.

[41] The code for filtering string forms in our natural-language processing experiments is included in Appendix E5.

[42] A major difference between the COMMON LISP and FRANZLISP versions of MORPH is that, due to space limitations on the smaller machines that run the COMMON LISP, it is necessary to use an index to facilitate access to the lexicon. The remarks here apply to the COMMON LISP version only.

Number of Total Entries:	11402
Bound Morphemes:	301
Phrasal Terms:	3321
Nouns:	7726
Verbs:	968
Adjectives:	2207
Adverbs:	101
Prepositions:	53
Other Entries:	347
Terms In TERMS+CAT:	6249
Mesh Terms:	3248

Figure 3S: Statistics on Composition of the Lexicon

Output. The output of MORPH is a list of feature lists of possible stems of the input word, in a form ready to be processed by the parser, RULEPAR-II.

2.3.4. How MORPH **Works**

In this section, we will first discuss some of the major functions in MORPH, then we will give typical examples to illustrate how MORPH works.[43]

In general, MORPH does three things. First, it attempts to match the input word with entries in the lexicon.[44] If a direct match is found, the complete lexical entry for the word is accessed. If a direct match is not found, MORPH performs a morphological analysis on the input word to determine if it is a morphological variant of an entry in the lexicon. Second, the feature list associated with the entry corresponding to that word is accessed. Third, once we have the feature list of the entry, grammatical information, derived from the category and morphological form of the word, is added to the feature list of the lexical entry and sent to the parser.

In the FRANZLISP version of the MORPH, the lexical entries are stored directly in memory, along with their feature lists as property values. Items can be accessed without searching. When given a word, MORPH first attempts to match the word with a lexical entry directly by testing whether it has a property value. If it does, this means there is a direct match and the feature list is returned. If it does not have a property value, this means the word is not included in the lexicon as a canonical term and it will require morphological analysis

[43] For more detail on the operation of MORPH, see Appendices E3. and E4., for both comments and source-code.

[44] **In its COMMON LISP version, it attempts to match words to the *index*.**

```
(anodmia 44154)
(anomaly 44192)
(anomia 44241)
(anopsia 44294)
(anorectic 44327)
(anorexia 44353)
(anorexia-nervosa 44535)
(anorexic 44597)
(anorexic 44644)
(anorthographic 44689)
(anorthographical 44720)
(anorthography 44753)
(anosmia 44801)
(anosmic 45000)
(anosphrasia 45024)
(anosphrasic 45066)
(another 45094)
(anoxia 45133)
(antacid 45181)
(ante= 45328)
(antecurvature 45353)
(anterior 45400)
(anterior 45515)
(anterior-leaflet 45565)
(anterior-pituitary 45613)
(anterior-pituitary-disease 45657)
(anterior-uveitis 45709)
(anteroposterior 45762)
```

Figure 39: Index to the Lexicon Used in COMMON LISP Version of MORPH

s er ed 's s^J y ly nd rd st th ing n't ally

Figure 40: The Suffixes Recognized by MORPH

to determine whether it is an inflected or derivational form of some entry or entries in the lexicon.

In the COMMON LISP version of MORPH, an index of the lexicon is loaded into the memory and stored in a hashtable. MORPH attempts to match an input w^{T}ord with terms in the lexicon by accessing the index first. If the word is in the index, (ż.e., the hashtable), the word has a corresponding lexical entry. The file position of its feature list is returned and the disk is accessed to load the feature list. If the word is not in the index, it is not in the lexicon. Morphological analysis is necessary to decided if it is a morphological variant of some lexical entry or entries in the lexicon. The function `find-stem` in both the FRANZLISP and COMMON LISP versions MORPH is used to do this matching.[45]

Once a direct match is attempted, even if successful, MORPH starts to perform morphological analysis to determine w$^?$hether the word is a morphological variant of some entry or entries in the lexicon. Note that some words might be both an entry in the lexicon and also a variant of some other lexical entry. For example, the word *alluring* is listed as an adjective in the lexicon. But it is also the present participle of the verb *allure.* We want MORPH to return both feature lists. The actual category of the word in use, whether it is an adjective or the present participle of the verb *allure,* will be decided by the parser after the parser gets the output from MORPH.

At present, MORPH recognizes just the suffixes given in Figure 40. These were chosen they are the most frequently occurring ones in the current lexicon. They represent suffixes for

- verbs in their third-person singular present form *(s);*
- verbs in their present-participle form *(-ing);*
- verbs in past- or past-participle form (-erf);
- nouns in their plural form *(s);* and
- adverbs derived from adjectives (-?/, -/?/, *-ally).*

MORPH starts to analyze a word by matching the ending of the word with the list of suffixes. The function `findsufs` is called inside the function `remsufs` to see if the ending of the word is a suffix. Taking *alluring* an example, again,, the suffix would be recognized as *-ing* and stripped off. At this point, the function `remsufs` calls the function `find-stem` to check if *allur* is in the lexicon—and fails, since *allur* is not a lexical entry. MORPH then starts the `stemtest` function to test the stem for well-formedness after its suffix has been

[45] The LISP code for the functions described in this section, along with comments, can be found in Appendices E3. and E4.

removed. In our example, after *-ing* is stripped off, stemtest makes sure that the stem does not end with *-e* or that the penultimate letter is a vowel. Thus, it would reject *allureing* as ill-formed while accepting, for example, *seeing*, though both end with *-e* after *-ing* has been removed. Other modifications of the stem are attempted in stemtest—such as adding *-e* after a *-y* is stripped off, as in the analysis of *countably*—but in the case of *alluring*, once the test on the stem is successful, find-stem is called again to see if there is a match with entries in lexicon. In this instance, since no modifications to the stem had been made, *allur* would again be determined to be a non-entry. MORPH then calls the function stemmods which modifies stems. In our example, *-e* is added to *allur* and find-stem is called again to match *allure* against lexical entries. This time it would be found as a lexical entry and its features list would be returned.

A description of the principal functions in MORPH is given in Figure 41.

2.3.5. Some Representative Examples

Consider several additional examples. Suppose the input contains the word *ache*. *Ache* is an entry in the lexicon, so find-stem matches the word with the entry in the lexicon directly. MORPH then loads the features associated with that entry from the lexicon, as shown in Figure 42. Since it is a verb and no morphological change has occurred, the features *form (pres inf)* and *png (s1 2 p1 p3)* are added to the feature list, as shown in Figure 43. This is the feature list for *ache* that is sent to the parser by MORPH.

Now consider the case of input word *aches*. *Aches* is not in the lexicon, so find-stem fails to find a direct match for the word. remsufs starts a morphological analysis of the word. find-sufs finds the suffix *s* and removes it. After testing by stemtest, which it passes, find-stem is called again and a match with the entry *ache* is found. The feature list associated with *ache* is loaded from the lexicon. Since the category of the lexical entry is *verb*, the word is recognized as the third-person present singular form of the verb *ache*. No stem modification is necessary since an entry was found. augment is then called to check that the category and the suffix agree—which they do. augment also adds more features to the feature list of *ache* and represents the word *aches* as shown in Figure 44. This is the feature list for *aches* that is sent to the parser.

In many cases, the morphological modifications are not as simple as the preceding example. Consider the word *identifies*. find-stem cannot find a direct match in the lexicon, but remsufs will be able to identify the ending of the word, *s*, as a suffix—so *s* is stripped off. Since *identifie* passes the stem test, find-stem is called and a match is attempted against lexical items. The match fails, indicating that some modification of the word may be necessary. Since *identifie* ends with *e*, stemmods calls another function (edrop) to remove the *e*; and the final *i* in *identifi* is swapped with *y*. The result is *identify*. find-stem is called again and this time succeeds. The feature list associated with *identify*, given in Figure 45 is returned. Since *identify* is a verb, *identifies* is recognized as its present third-person singular form. So its feature list is augmented as shown in Figure 46.

Consider as a final example a more detailed description of MORPH's handling of *alluring*, anaw Stands for 'analyze-word'. Takes the input, attempts a direct match, and starts the morphological analysis. It takes a word as its input and returns a list or lists of feature lists if the word has a direct match or is a morphological variant of some entry in the lexicon; it returns *nil* otherwise.

find-stem Tries to match a word—either the input word or the input word with suffix(es) stripped off—to a lexical entry. It is called when the first direct match is attempted: also after some morphological modifications have been made to the input word. It returns the feature list of the lexical entry as recorded in the lexicon if there is a match between the word and the lexical entry; Otherwise, returns *nil*

remsufs Stands for 'remove-suffix'. It initiates morphological analysis. It calls findsufs to match the ending of a word with the list of suffixes. Once there is a match, remsufs will strip off the suffix, remsufs also calls stemtest, find-stem, and stemmods.

findsufs Matches terminal characters in input words against the list of known suffixes.

stemtest Tests the well-formedness of the portion of the word remaining after its suffix has been removed. For example, it makes sure MORPH accepts *alluring* but not **allureing.* If stemtest does not return *nil*—which means the well-formedness test has been passed—find-stem will be called by remsufs to match the word against lexical entries.

stemmods Stands for 'stem-modifications'. In case stemtest returns *nil,* no attempt will be made to match the word with a lexical entry immediately. Instead, stemmods is called to modify the ending of the word. It also makes sure that in words like *bigger,* the stem-final consonant is degeminated. After stem modifications, find-stem will be called to see if there is a match. It there is a match, the feature list will be loaded. If there is no match, it means that the word is not a morphological variant of some entry in the lexicon, and *nil* will be returned.

augment Checks whether the suffix that has been stripped off agrees with the category of the entry returned by find-stem. If they do not agree, the word is rejected as a morphological variant of the lexical entry. If the suffix and the category agree, augment also adds grammatical information to the feature list of the entry. This grammatical information is derived from the syntactic category of the input word, which is either listed in the lexicon or determined by the morphological analysis, and includes values for *category*, *form*, *png*, *case*, *etc.*

Figure 41: Descriptions of Functions in MORPH

(ache cat v subcat intrans **syn** pain)

Figure 42: Lexical Entry for *Ache*

```
(sense ache cat v subcat intrans syn pain
                    form (pres inf) png (si 2 pi p3))
```

Figure 43: The MORPH-Analyzed Lexical Entry for *Ache*

```
(sense aches cat v subcat intrans syn pain suffix s
                    form pres png (sm sf sn))
```

Figure 44: The MORPH-Analyzed Lexical Entry for *Aches*

characterized above. *Alluring* is listed as an adjective in the lexicon, so the first attempt by find-stem is successful and the feature list is loaded, remsufs attempts morphological modification of the word. The suffix is recognized to be -*ing* and it is stripped off. At this point, the function remsufs calls the function find-stem to check if *allur* is in the lexicon—and fails since *allur* is not an entry. The function stemtest is called to test the well-formedness of the word after its suffix has been removed. In our case, stemtest makes sure that after -*ing* is removed, either the word does not end with -e or the penultimate letter is a vowel. Since no match has been found after stemtest, the function stemmods is called, *e* is added and find-stem is called again to match *allure* against lexical entries. An item is found and its feature list is loaded. Augment checks both of the feature lists to check that their categories and suffixes agree. In this case they do, so after augmenting, the feature lists are returned as shown in Figure 47.

```
(identify cat v subcat trans)
```

Figure 45: Feature List of *Identify*

```
(sense identify cat v subcat trans suffix s form pres png (sm sf sn))
```

Figure 46: The MORPH-Analyzed Lexical Entry for *Identifies*

```
(sense alluring cat adj syn seductive form pos)

(sense allure cat v subcat trans suffix ing form prog)
```

Figure 47: The MORPH-Analyzed Lexical Entries for *Alluring*

3. Parsing Natural Language

3.1. Introduction

The MedSORT-II Project has developed both a large frame-based thesaurus of medical concepts and a program, SPAWN, to manage the thesaurus. We have adapted an existing natural-language processing program to demonstrate the correctness of SPAWN and the thesaurus. Specifically, the case-frame parser RULEPAR, implemented in a FRANZLISP-based rule language and agenda structure called RULEKIT, has been adapted to parse English sentences corresponding to medical laboratory-source observations, producing semantic representations in the form of instantiated frames. We call our updated version of RULEPAR "RULEPAR-II".[46]

The major objectives of this aspect of the Project are twofold. First, to demonstrate the quality of the lexicon and thesaurus: that appropriate meanings exist for entries and that the entries are compatible with each other. Second, to show the utility of SPAWN, the system we have developed to manage semantic knowledge.

3.2. Procedure and Rationale

There are a number of considerations affecting decisions on natural-language processing, depending on the type of applications one plans. The goal of a natural-language interface is to translate input from the user—in this case, medical *laboratory-source* observations—into a semantic representation that can be dealt with by a computer. This translation of natural-language expressions into a semantic representations is known as *parsing*, and makes up a complex and interesting field of Artificial Intelligence.[47] The format of the semantic representation here is the case-frame representation described in the knowledge representation section of this report. As a result, the desired parser must use case-frames as its semantic representations.

For MedSORT-II, input expressions are read as character-strings and mapped to lexical items through the morphological mapping package, MORPH. For example, the word *eaten* would be mapped to the case frame for its root *eat*. When reading an input expression, hyphens may be substituted for spaces so that lexical entries corresponding to more than one English word will be recognized. For example, the input string *amyl nitrite* will match against the lexical entry *amyl-nitrite*. As a default, it has been useful to consider only the *longest* entries in the lexicon that are obtained from a given input string.

Once it has been decided that the underlying semantic representations are case-frames, there is essentially one other decision: whether the parser should be *top-down* or *bottom-up*. A strictly bottom-up parser assigns structure only to input it has seen; a top-down parser hypothesizes structures that get filled (or discarded) as the input is parsed. For example, suppose the following rule is part of a given grammar:

[46] We include the source code for RULEKIT in Appendix F1.; the source code for RULEPAR-II in Appendix F2.

[47] A good source for papers and background information on this field is [Gross, *et al.* 1986].

- *NounPhrase* —» *Determiner* *Noun*

Also suppose that the word *the* is entered in the lexicon under the category *Determiner.* Upon receiving the input *the,* a top-down parser will hypothesize that a noun is coming next (along perhaps with other hypotheses deriving from other phrase structure rules) while a bottom-up parser will make no hypothesis at all. It will merely know that a determiner has been encountered. If a noun appears next in the input stream, a noun phrase would be proposed by the bottom-up parser.

To help choose between top-down and bottom-up strategies, it is useful to note that the bottom-up strategy always returns some structure, as long as the input expressions are found in the lexicon. The parser-generated structure may not necessarily fit together as one frame, reflecting limitations in the grammar rules, lexicon, or knowledge base.. If a top-down parser encounters the same input, it will return nothing, since there is no complete parse. The bottom-up parser will at least return the pieces.

3.3. RULEPAR-Il: **A Lexically-Driven Case-Frame Parser**

Based on the above considerations, it has been decided to use RULEPAR-II, a lexically-driven bottom-up, case-frame parser, for MedSORT-II Project applications. RULEPAR-II is implemented in a FRANZLISP-based rule language and agenda structure called RULEKIT.[48]

There are a number of differences between the original RULEPAR and RULEPAR-II. Since the original RULEPAR was based on FRAMEKIT, the predecessor to SPAWN, changes were necessary in the function calls to the underlying frame-representation language. Secondly, RULEPAR provided only trivial morphological analysis, while RULEPAR-II calls MORPH, a sophositicated analyzer. Thirdly, and most importantly, the grammar of RULEPAR-II has been adapted the test domain, *medical laboratory-source observations.*

Since RULEPAR-II is lexically driven, it will always return at least a partial representation to the expressions it encounters, as long as some of the entered words are present in the lexicon. Although in an ideal state one would like to have a complete representation for each input expression, it will always be the case that some expressions may be analyzed as incomplete (by the available rules). In such cases it is desirable to return as much parsed structure as possible.

A parsed expression is built using complex case frames that are stored in the knowledge base. The form of such frames is given schematically in Figure 48. The slot *cases* contains a list of allowable case-slot names in the frame. At each of these case-slots, the facet *semantics* contains a disjunction of acceptable frame types. The facet *fr-markcr* contains the allowable frame markers, usually in the form of prepositions. The facet *slot-restriction* contains the name of a LISP function that is applied to a frame that is attempting to fill the case-slot. If this function returns some non-n?7 value, the new frame is added as the value of the case-slot.

As an example, consider the sample manifestation frame given in Figure 49. It represents the information that would be required to yield a well-formed, complete clinical finding in

[48] **The rule language was written by Jaime Carbonell under the MedsoRT-l Project. The parser is based on RULEPAR, also written by Carbonell (see Appendix D.3 of [Carbonell,e*** ***ai*** **1985]).**

```
(make-frame <frame-name>
   (isa (value <parent-framel> <parent-frame2> ...))
   (cases (value <casel> <case2> ...))
   (<casel> (semantics <semantic-class-restrictionl>
                       <semantic-class-restriction2> ...)
            (fr-marker <case-markerl> <case-marker2> ...)
    (slot-restriction <function-name>))
   (<case2>_____)
   ...)
```

Figure 48: Form of a Complex Frame in the Knowledge Base

```
[make-frame *indirect-imaging-mx
            (isa (value *imaging-mx))
            (cases (value *method *source *result))
            (method (semantics *lab-observation-technique-cx)
                    (fr-marker by on in with for of to))
            (source (semantics *body-thing-cx)
                    (fr-marker at on with by from in of on to via with)
                    (slot-restriction observation-mx-part-in-region-p))
            (result (semantics *pathological-factor-cx
                               *disease-cx
                               •measure-theoret ic-thing-cx))]
```

Figure 49: *Indirect-Imaging* Manifestation Frame

any observation involving the laboratory methods included under *indirect-imaging—x-ray.* for example. It shows, basically, that any such observation must contain information about the specific *method* used, the body area, or *source,* under observation, and the outcome or *result* of the procedure. A sample parse that uses this frame can be found in Figure 50. giving RULEPAR-H's rendering of the statement *chest x-ray of the lung revealed parenchynal calcification.* That statement should qualify as a finding; and the effect of the parse is to make explicit what role the various concepts play and what relations exist among them.

3.4. Parsing with RULEPAR-II

RULEPAR-II parses in the following manner. Words from the input stream are morphologically analyzed by **MORPH** and then matched against the lexicon. Combinations of consecutive words separated by hyphens are also compared with the lexicon to retrieve maximal-length lexical entries. An entry returned by MORPH has the form given in Figure 51.

```
[make-frame *mx76
    (isa
      (value *indirect-imaging-mx))
    (mx-rank
      (value 3))
    (source
      (value
             [make-frame *np490
                 (isa
                   (value *macro-body-part-cx))
                 (head
                   (value *lung))
             [make-frame *np487
                 (isa
                   (value *macro-body-region-cx))
                 (head
                   (value *chest))] ))
    (method
      (value
             [make-frame *np488
                 (isa
                   (value *indirect-imaging-cx))
                 (head
                   (value *xray))] ))
    (result
      (value
             [make-frame *np493
                 (isa
                   (value *pathological-action/process-cx))
                 (head
                   (value *calcification))
                 (loc
                   (value
                          [make-frame *np492
                              (isa
                                (value *macro-body-region-cx))
                              (head
                                (value *parenchymal))] ))] ))]
```

Figure 50: Parse: *Chest X-Ray of Lung Revealed Parenchymal Calcification*

```
(<word>
    cat <part-of-speech>    ; e.g., n, adj, prep,
    val <bx>                ; either bx or nil, depending on whether
                            ; entry is in the bx hierarchy
    sense <frame-name>      ; thesaurus pointer
    number <val>            ; e.g. singular, plural
    ref <val>               ; reference, e.g. definite, indefinite
                            ; this entry makes sense only for determiners
    quant <val>             ; e.g. 57, 1
    form <val>              ; form of verb, e.g., pastprt, present
   )
```

Figure 51: The Form of the Lexical Entries Returned by MORPH

MORPH actually returns quite a bit more information than shown in Figure 51, but the fields mentioned are the only ones used by RULEPAR-II. For example, MORPH has the ability to return many entries for any given word. RULEPAR-II, however, is naive in some respects and can only use one entry per word. A simple ordering of syntactic categories allows RULEPAR-II to select among entries returned from a word. Typical expressions in the domain of medical observations contain more information in noun-phrases than in verbs. Thus, almost all of the semantic content of a given observation comes from its syntactic nouns and adjectives; little from its verbs. As a result, if an entry contains a syntactic noun and verb, only the noun is passed on to RULEPAR-II.[49]

Figure 51 shows some of the *features* and *values* derived from the Lexicon and MORPH. that are utilized in RULEPAR-II processing. These are described below.

- The field *cat* indicates syntactic category and is used in the syntactic rules of composition defined in the grammar of RULEPAR-II.
- The field *val* indicates whether or not the entry corresponds to a frame in the basic (Bx) hierarchy. If so, the value *bx* apears there. If this field is empty, the word will not have a frame associated with it in the parse of a phrase containing it. It may, however, contribute to other frames that are built from the input.
- The field *sense* is the name of the actual frame in hierarchy.
- The field *number* indicates whether or not the item is singular or plural.
- The field *ref* gives the entry's reference, either definite or indefinite. It is only defined on the category *determiner.*

[49] This restriction in parsing strategy reflects a conscious design decision, built into RULEPAR-II. Under modification, to allow for parallelism and back-tracking in processing, RULEPAR-II would be capable, in theory, of processing arbitrary syntactic structures. Very likely, however, it would prove more efficient to use a different style of parsing than the one offered in RULEPAR-II, if one wished to capture a wider range of syntactic phenomena.

- The field *quant* gives a numeric value to the entry. This field is only defined for items in category *numeral*
- The fields *number, ref* and *quant* are used to copy their information to the resulting frames.
- The field *form* indicates a verb's form, *e.g., past-participle.* In RULEPAR-II, this field is used in the recognizing of passives. In a more syntactic parser this field would have more use.

Once morphological analysis has taken place, syntactic rules of composition are then used to propose larger syntactic units *{e.g.,* noun-phrase *(np)* or prepositional phrase *(pp))-* The proposed larger syntactic units are then tested against semantic information obtained from the MedSORT-II Thesaurus. If compatible, the frame representation is returned. To paraphrase the characterization of this process in [Carbonell, *et al.* 1985], "Syntax *proposes a structure^* semantic information *certifies* it."

The RULEPAR-II grammar presently handles the following syntactic phenomena.

1. *Noun-noun compounds* (any number of nouns). Nouns are attached to following nouns based on semantics and cross-slot constraints. The attachment procedure is deterministic, with preferential treatment given to local attachment. The semantics is given by the external SPAWN knowledge base.
2. Any number of *adjectives before a noun.* The attachment procedure is the same as with noun-noun compounds.
3. *Determiners* and *numerals* at the start of a noun phrase.
4. *Prepositional phrases,* and *prepositional-phrase attachment to a noun phrase,* possibly inside another prepositional phrase. Any number of attachments are possible. This attachment procedure is semantically-driven, governed again by adjacency preferences and crossing constraints. Permissible prepositions for case attachment are listed in the *fr-marker* facet of the *case* slot in thesaurus entries, and are inherited.
5. *Verb Phrases. Subjects* and *direct objects* are attached to the verb frame, subject to the semantic restrictions in the frame. Permissible prepositions for case attachment are listed in the *fr-marker* facet of the case in thesaurus entries, and are inherited. The parser also handles *active* and *passive* by listing the slots compatible with an active subject in the *subj* slot, *active* facet; and listing the slots compatible with a passive subject in the *subj* slot, *passive* facet. Direct objects are handled in a similar manner with the slot *dircct-obj.* Verbs with semantic content are not prevalent in medical observations, however, so this syntactic phenomenon is not exploited in our applications.

Once syntactic processing has been carried out on a given laboratory-source observation, the grammar attempts to assemble a manifestation frame representing the entire observation. This assembling is dependent on the structure of the thesaurus stored in the frame organizer, SPAWN. The Thesaurus consists of three conceptually distinct hierarchies with limited, well-defined connections among them. The three hierarchies are known as the *Bx* (Basic), *Cx*

```
(make-frame <bx-frame>
   (header-for-cx (value <cx-frame>))...)
```

Figure 52: Schematic Structure of a Bx-Frame/Cx-Frame Link

```
  (make-frame <cx-frame>
     (cx-header (value <mx-frame>))...)
```

Figure 53: Schematic Structure of a Cx-Frame/Mx-Frame Link

(Complex) and *Mx* (Manifestation) Hierarchies, as described in Part II of this report. Bx frames typically have no slots associated with them besides ***isa*** links, which run up into the Bx hierarchy, and ***header-for-cx*** links which hook up to the Cx hierarchy. Bx frames provide the basic structures from which complex frames are built. The ***header-for-cx*** slot is used by RULEPAR-II to build complex frames efficiently. Since a complex frame needs a head (by definition), it is very efficient to build the Cx frames by traversing this head.

In Figure 52, <*bx-frame*> is the head for <*cx-frame*>. For example, the Bx frame for a body part such as ***arm*** might link up to the Cx frame ***macro-body-part-cx*** through ***macro-body-part-bx***. In the ***macro-body-part-cx*** frame there may be slots corresponding to ***quality***, ***index***, ***etc.*** The link between the two hierarchies (in this case between ***macro-body-part-bx*** and ***macro-body-part-cx***) is the slot called ***header-for-cx*** in this parser.

Mx frames are frames that represent pragmatic units in the given domain. In the domain presented, medical *laboratory-source* observations, *tissue-technique-mx* could be one such Mx frame. Mx frames have slots that are pragmatically based, reflecting the privileged constellations of concepts that constitute domain knowledge. For example *indirect-imaging-mx* has slots corresponding to *source*, *result* and *method*, since medical source observations are desired. In RULEPAR-II, there are links running from the Cx to Mx hierarchies called *header-for-mx*. Once again, these links enable efficient parsing. Mx frames are only built if a Cx head already exists. A schematic example of such links is given in Figure 53. There, <*cx-frame*> is the head for <*mx-frame*>.

RULEPAR-II also has the ability to rank Mx frames such that if more than one is triggered by the input, only the higher priority frame will be generated. Such rankings are clearly domain dependent. For example, in medical *laboratory-source* observations, laboratory ***assay*** procedures have higher priority than *extraction* procedures, from the point of view of information granularity. So if ***biopsy***, an extraction procedure, occurs in the same sentence as ***stain***, an assay procedure, the Mx frame ***assay-procedure-mx*** is triggered. In our test knowledge base, the frame for ***assay-procedure-mx*** has a ***method*** slot which is compatible with any ***laboratory-source*** procedure. In a sentence that contains both ***biopsy*** and ***stain***, the Cx frames for each would fill this slot. The stain is interpreted as the more important method since it triggered the Mx frame. The general form for a ranked Mx frame is given

```
(make-frame <mx-frame>
   (mx-rank (value <number>))...)
```

Figure 54: Schematic Form of Imposed Rankings of Mx Frames

```
(make-frame <bx-frame>
   (def-trigger-for-cx (value <cx-frame>))...)
```

Figure 55: Schematic Form of a Default Trigger in Bx Frames

in Figure 54.

If no Mx frame is triggered by a Cx frame in a given parse then all the Bx frames found by RULEPAR-II are checked for a default trigger through the slot *def-trigger-for-cx.* The value of this slot is another Cx frame that will be the head for an Mx frame. As a result Mx frames can be triggered without mentioning the formal head of the frame. For example, the input *light chains in urine have increased* has no explicit method mentioned; it is inferred to be *immunoelectrophoresis.* This default is found in the Bx frame for *light-chain* since in the context of medical *laboratory-source* observations light chains are nearly always the object of the *substance-technique* immunoelectrophoresis. In the schematic frame given in Figure 55, *<bx-frame>* is a default trigger for *<cx-frame>,* where *<cx-frame>* is the head for some Mx frame.

Default slots on Bx frames are also available to represent complex concepts that appear as a single word, such as *bronchoscopy,* which is a *direct-imaging* procedure that has the location *bronchial* built in. Having a Bx frame for the concept *bronchoscopy* is not optimal: *bronchoscopy* is actually a complex concept, composed of a *method* and *location.* It would be theoretically more sound to perform morphological analysis on the word *bronchoscopy.* to decompose it into the bound morphemes *bronch=.* the *location,* and *=oscopy.* the *procedure.* each represented by separate Bx-level frames. We have not yet, however, implemented such a facility in MORPH; so our Thesaurus maintains this style of sub-optimal representation. To effect the appropriate decomposition of information—making explicit the entailed *location* of *bronchoscopy,* for example—we utilize *default* slots on Bx frames. The general form for default slots in a Bx frame is shown in Figure 56. There, the values stored in *default-slots* are the slots of the *<bx-frame>* that are needed to make up a complex of concepts associated with the *<bx-frame>.*

3.5. Experiments in Processing Clinical Findings

Using RULEPAR-II, we have experimented with the parsing of natural-language statements designed to express clinical findings in the domain of *laboratory-source observations.* For our demonstration, we used the corpus of statements collected in Appendix F3., a sample of

```
(make-frame <bx-frame>
   (default-slots (value <slotl> <slotl> ...))
   (<slotl> (value <valuel>))
   (<slot2> (value <value2>))
   ...)
```

Figure 56: Schematic Form of Default Slots in Bx Frames

```
Sentence 1:  (ameba gel diffusion was positive)
Sentence 2:  (csf was obtained and animal inoculation revealed toxoplasma)
Sentence 3:  (legionella direct fluorescent antibody on sputum came out positive)
Sentence 4:  (light chains in urine have increased)
Sentence 5:  (positive on ascorbic acid loading test)
Sentence 6:  (mycoplasma complement fixation test was performed and was positive)
Sentence 7:  (radioimmunoassay on urine for legionella antigen is positive)
Sentence 8:  (lepromin skin test was positive)
Sentence 9:  (acid fast bacteria found with csf smear)
Sentence 10: (a transbronchial biopsy was performed and Candida was detected by stain)
Sentence 11: (open biopsy of lung revealed aspergillus by stain)
Sentence 12: (serum vdrl was positive)
Sentence 13: (no specific change in eeg)
Sentence 14: (eeg showed multiple spike burst)
Sentence 15: (atrial flutter found during ekg)
Sentence 16: (first degree heart block detected with ekg)
Sentence 17: (an ekg was performed and showed a short pr-interval)
Sentence 18: (tachycardia in sinus detected during ekg)
Sentence 19: (flat t-wave on ekg)
Sentence 20: (repetitive high frequency potential with needle insertion during emg)
```

Figure 57: Part of Corpus of Parsed Statements

which is given in Figure 57.

Results of our parses are given in Appendix F4. Typical parses are given in Figures 50 and 58. The grammar we used is given in Appendix F5.

We believe these experiments demonstrate not only the soundness of the design of the MedSORT-II Thesaurus and knowledge-base management system. SPAWN, but also the feasibility of developing general natural-language processing facilities in the domain of clinical biomedicine. A more complete demonstration of the promise of these results awaits further development of the MedSORT-II Thesaurus and the grammar of RULEPAR-II.

```
Sentence 2: (csf was obtained and animal inoculation revealed toxoplasma)

104 rules tested, 17 rules fired in 5 seconds CPU.

[make-frame *mx6
    (isa
      (value *substance-technique-mx))
    (type
      (value mx))
    (mx-rank
      (value 1))
    (source
      (value *np26))
    (method
      (value *np28))
    (result
      (value *np30))]
----------------------mx6----------------------
[make-frame *mx6
    (isa
      (value *substance-technique-mx))
    (type
      (value mx))
    (mx-rank
      (value 1))
    (source
      (value
            [make-frame *np26
                (isa
                  (value *body-fluid-cx))
                (head
                  (value *csf))
                (quantifier
                  (value singular))
                (type
                  (value np))
                (parent
                  (value *mx6))))
    (method
      (value
            [make-frame *np28
                (isa
                  (value *substance-technique-cx))
                (head
                  (value *animal-inoculation))
                (quantifier
                  (value singular))
                (type
                  (value np))
                (parent
                  (value *mx6))))
    (result
      (value
            [make-frame *np30
                (isa
                  (value *protozoa-cx))
                (head
                  (value *toxoplasma))
                (quantifier
                  (value singular))
                (type
                  (value np))
                (parent
                  (value *mx6))))
```

Figure 58: Parse: *Csf was Obtained and Animal Inoculation Revealed Toxoplasma*

Part IV

References

Brachman 1979 R.J. Brachman, On the epistemological status of semantic networks, in **N.V. Findler (Ed.),** ***Associative Networks: Representation and Use of Knowledge by*** *Computers,* New York: Academic Press, 1979. 3-50.

Brachman &: **Levesque 1985** R.J. Brachman and H.J. Levesque (Eds.), *Readings in Knowledge Representation,* Los Altos, CA: Morgan Kaufmann Publishers, Inc., 1985.

Carbonell *$z* Hayes 1981 J.G. Carbonell and P.J. Hayes, Dynamic strategy selection in **flexible parsing, in** ***Proceedings of the 19th Annual Meeting of the Association for Computational*** *Linguistics,* The Association for Computational Linguistics, 1981, 143-147.

Carbonell & Hayes 1983 J.G. Carbonell and P.J. Hayes, Robust multi-strategy parsing. Technical Report, Department of Computer Science, Caxnegie Mellon University, 1983.

Carbonell, *et al.* **1983** J.G. Carbonell, W.M. Boggs, M.L. Mauldin, and P.G. Anick, The XCALIBUR Project, a natural language interface to expert systems. Technical Report, Department of Computer Science, Carnegie Mellon University, 1983.

Carbonell, *et al.* **1986** J.G. Carbonell, D.A. Evans, D.S. Scott, and R.H. Thomason, On the design of biomedical knowledge bases, in R. Salamon, B. Blum, and M. Jorgensen **(Eds.),** ***Medinfo 86: Proceedings of the Fifth Conference on Medical Informatics,*** **Amsterdam:** Elsevier Science Publishers, 1986, 37-41.

Carbonell &: Thomason **1986** J.G. Carbonell and R.H. Thomason, Parsing in biomedical indexing and retrieval, in A.H. Levy and B.T. Williams (Eds.), *Proceedings of the AAMSI Congress 86,* Anaheim, California. May 8-10. 1986, 274-277.

Dorland 1985 *Dorland's Illustrated Medical Dictionary.* 26th Edition, Philadelphia. PA: W.B. Saunders. 19S5.

Evans &; Miller 1987 D.A. Evans and R.A. Miller, *Initial Phase in Developing Representations* ***for Mapping Medical Knowledge:*** **INTERNIST-1/QMR, HELP,** ***and*** **MESH,** Unified Medical Language System (UMLS) Project Report, May 10, 1987. (Available as Technical Report CMU-LCL-87-1, Laboratory for Computational Linguistics, Carnegie Mellon University.)

Evans &; **Scott 1986** D.A. Evans and D.S. Scott, Concepts as Procedures, to appear in ***ESCOL-86, Proceedings of the Eastern States Conference on Lingusitics,*** **Ohio University** Press.

Grishman & Kittredge 1986 R. Grishman and R. Kittredge (Eds.), *Analyzing Language* ***in Restricted Domains: Sublanguage Description and Processing,*** **Hillsdale, NJ: Lawrence** Erlbaum, 1986.

Gross, *et al* **1986** B.J. Gross, K.S. Jones, and B.L. Webber, *Readings in Natural Language Processing,* Los Altos, CA: Morgan Kaufmann Publishers, Inc., 1986.

Horty, *et al* **1986** J. Horty, R.H. Thomason, and D.S. Touretzky,, *A Skeptical Theory of Inheritance in Nonmonotonic Semantic Nets.* Technical Report, Department of Computer Science, Carnegie Mellon University, 1986.

Levesque &: **Brachman 1985** H.J. Levesque and R.J. Brachman, A fundamental tradeoff in knowledge representation and reasoning,, in R.J. Brachman and H.J. Levesque (Eds.), *Readings in Knowledge Representation,* Los Altos, CA: Morgan Kaufmann Publishers, Inc., 1985, 42-70.

Miller, *et* ***al*** **1982** R.A. Miller, H.E. Pople, and J.D. Myers, INTERNIST-I, an experimental computer-based diagnostic consultant for general internal medicine. *New England Journal of Medicine,* Volume 307, 1982, 468-476.

Miller, *et* ***al*** **1986a** R.A. Miller, F.E. Masarie, and J.D. Myers, Quick medical reference for diagnostic assistance. *AID Computing,* Volume 3, 1986, 34-48.

Miller, *et al.* **1986b** R.A. Miller, M.A. McNeil, S. Callinor, F.E. Masarie, and J.D. Myers, Status report: The INTERNIST-l/Quick Medical Reference Project. *Western Journal of Medicine,* December, 1986.

Sager, *et* ***al*** **1987** N. Sager, C. Friedman, and M.S. Lyman, *Medical Language Processing: Computer Management of Narrative Data,* Reading, MA: Addison-Wesley, 1987.

Steele 1984 G.L. Steele, Jr., *Common Lisp: The Language,* Bedford, MA: Digital Press, 1984.

Taber 1981 *Taber's Cyclopedic Medical Dictionary* C.L. Thomas (Ed.), Philadelphia, PA: F.A. Davis, 1985.

Thomason, *et al* 1986 R.H. Thomason, J. Horty and D.S. Touretzky, ,4 *Calculus for Inheritance in Monotonic Semantic Nets.* Technical Report CMU-CS-S6-1S3. Department of Computer Science, Carnegie Mellon University. 19S6.

Touretzky 1984 D.S. Touretzky, *LISP: A Gentle Introduction to Symbolic Computation,* New York: Harper and Row, Publishers, 1986.

Touretzky 1986 D.S. Touretzky, *The Mathematics of Inheritance Systems.* Los Altos. CA: Morgan Kaufmann Publishers, Inc.. 1986.

Touretzky, *et al* **1986** D.S. Touretzky, J. Horty, and R.H. Thomason, *Issues in the Design of Nonmonotonic Inheritance Systems.* Technical Report, Department of Computer Science, Carnegie Mellon University, 1986.

Vries, *et* ***al*** **1986** J.K. Vries, P. Shoval, D.A. Evans, J. Moossy, G. Banks, R. Satchaw, An expert system for indexing and retrieving medical information, August 16, 1986.

Woods 1975 W.J. Woods, What's in a link: Foundations for semantic networks, in D.G. Bobrow and A.M. Collins (Eds.). *Representation and Understanding: Issues in Cognitive Science* , New York: Academic Press, 1975, 35-82.

Part V

Guide to Appendices

The following outline of the structure and contents of the Appendices is designed to provide a guide to the MedSORT-II-Project accompanying documentation.

A Project Background Documents

Al. Statement of Proposed MedSORT-II Activity

This appendix contains a portion of the original proposal for MedSORT-II Project activity.

A2. Slides of 6-Month Report

This appendix contains copies of the slides used in the presentation of the 6-month report on the MedSORT-II Project at the NLM, January 16, 1987.

B Thesaurus/Knowledge Base Documents

Bl. Final Report on UMLS Task 2

This appendix contains a copy of Part II of the UMLS Task-2 final report (and appropriate appendices).

B2. Internist-I/QMR Terms in Context

This appendix contains a sample of the index of IXTERNIST-l/QMR terms and their associated findings. The original file is 31.576 lines long.

B3. The Basic (Bx) Hierarchy

This appendix contains the basic hierarchy, originally developed under Task 2 of the Unified Medical Language System (UMLS) Project at Carnegie Mellon University.

B4. Classification Index

This appendix contains a sample set of terms the terms in the thesaurus, indexed according to their semantic category. (Note: The **full** index can be found in Appendix Bll.)

B5. The Manifestation (Mx) Hierarchy

This appendix contains the manifestation (Mx) hierarchy, augmented with categories derived from the MedSORT-I thesaurus.

B6. The Thesaurus Hierarchy, Augmented for MedSORT-I Concepts

This appendix contains the basic (Bx) hierarchy, augmented with categories contained in the MedSORT-I thesaurus.

B7. Internist-I/QMR Findings with Their Semantic Classifications

This appendix contains the INTERNIST-l/QMR findings, classified according to the Manifestation (Mx) Hierarchy. Each finding is paired with its representation in terms of semantic categories from the Basic (Bx) Hierarchy. The files were generated using an earlier version of the terms classification, and therefore do not reflect some of our more recent categories and revisions.

B8. Mx and Cx Frames

This appendix contains the frame-based grammar developed for use in natural-language processing with RULEPAR-II. It differs somew^{T}hat from the grammar utilized in the system-demonstration program. There are 15 Mx frames and 173 Cx frames.

B9. Test Phrases

This appendix contains sample input used to test RULEPAR-II and the frame-based grammar. The phrases are arranged according to the Mx frame that they should instantiate. This set of examples differs slightly from the set actually used in the Project demonstration program.

BIO. Listing of MedSORT-II Thesaurus Bx Classification of Terms

This is a listing of all the terms that appear in the MedSORT-II thesaurus/knowledge base, classified according to Bx-hierarchy category. The sources for the terms found in this listing include the INTERNIST-l/QMR listing of findings, the MARS Thesaurus, and the MedSORT-I Thesaurus. Some of the more esoteric categories came from the MedSORT-I classification scheme that was built into the MedSORT-II Bx hierarchy. Refinements of several categories reflecting terms that appear in *lab-assay-procedure, imaging-,* and *extraction- procedure* findings were made from information derived in the analysis of the files listed in B16. For a listing of these terms indexed by their Bx classification, see Appendix B1l.

Bll. Index of the MedSORT-II Thesaurus Terms

This index lists every classification category in the MedSORT-II Bx hierarchy alphabetically followed by all terms that had a classification placing them in that category. This file was generated from the information contained in Appendix BIO. Many terms are multiply classified. To see where each category fits into the MedSORT-II hierarchy, see Appendix B6. Categories beginning with 'direct-imaging', 'indirect-imaging', 'substance-technique', 'tissue-technique', 'fluid-extraction', and 'tissue-extraction' represent refinements of the larger categories in the hierarchy to delimit terminology that appears only in the context of these particular lab-procedures in the iNTERNIST-l/QMR findings. *(Cf,* also, Appendix B16.)

B12. Listing of Entries Given a Significant *Pathology-Class* in the MARS Thesaurus

This is a full listing of all those entries in the MARS Thesaurus that were given a 'significant' (by our estimation) *pathology-class* field. Those *pathology-class* entries we considered significant were: *disease, anatomy, noun, adjective, procedure, manifestation, substance, physiology, animal,* and *etiology.* Note that many of these entries are long multi-word phrases that would be better broken down into smaller conceptual units. This was done for all categories except *substance,* for which all terms not already found in the thesaurus were entered before a study could be made of which constituents among those terms could be considered an atom. In total, 713 of these entries were already found in the MedSORT-II Thesaurus as of May 13. For a listing of all the terms we thought should be entered from this list, see Appendix B13. For a listing of those we had time to classify and enter, see Appendix B14.

B13. Listing of All Atomic Concepts Drawn from the MARS Thesaurus

This appendix is a listing of all the concepts found in the MARS Thesaurus deemed to be atomic. We found these terms by examining MARS entries up to three words in length, choosing only those that represented an atomic concept, and then checking to see what words in Appendix B12. were not found either among those terms or in the MedSORT-II Thesaurus of May 13. At the time of creation, none of the terms in this list appeared in the MedSORT-II Thesaurus. Because of our method of finding these terms, all the entries in Appendix B12. can be constructed from this list and the May 13 edition of the MedSORT-II Thesaurus. Not all could be added to the thesaurus due to limitations in time and resources. For a listing of those terms that were entered, see Appendix B14.

B14. Listing of MARS Thesaurus Terms Entered in the MedSORT-II Thesaurus

This is a listing of the terms from the MARS Thesaurus that were classified according to the MedSORT-II Bx hierarchy and added to the knowledge base. All terms drawn from a MARS Thesaurus entry given the *pathology-class disease, procedure, manifestation, substance, physiology, animal,* and *etiology* were added. The terms drawn from a MARS entry with a *pathology-class* of *anatomy, adjective,* and *noun,* as well as the list of terms not in our preliminary set of atomic concepts, were not classified due to limits on time and resources.

This means that approximately 45% of our projected addition from the MARS Thesaurus was completed, adding almost 1000 new terms.

B15. Listing of Terms with Special Relations in the MedSORT-II Thesaurus

This is a listing in SPAWN-readible format of terms that have additional information added by specially defined relations. These relations are listed first so that SPAWN can build the network of semantic links correctly. The *add* command, which has the syntax "*add <term-1> <relation> <term-2> ... <term-n>*" simply tells SPAWN to link *<term-1>* to *<term-2> ... <term-n>* by a link of type *<relation>*. This file, in conjunction with Appendix B10. forms the whole of the MedSORT-II knowledge base.

B16. Analysis Files Used in Subclassifying of Lab Procedure Terminology

This appendix consists of three files that were used in generating subclassifications for terms appearing in *lab-procedure* findings. Each file was generated by examining the listing of classified findings and decomposing into MedSORT-II terms those that contained a *lab-procedure* term of the appropriate sort. The terms are sorted by their Bx-hierarchy classification, and include an analysis of how often they occurred in each of two types of findings. The first file is an comparison of terms in findings with either a *tissue-technique* or *substance-technique* term. The next compares terms from the classes *tissue-extraction* and *fluid-extraction.* The last compares terms in *indirect-* and *direct-imaging* findings. Only certain classes from each of these listings was chosen to be subclassified. See Appendix B11. for a list of subclasses that were actually created and for their classified terms.

B17. MedSORT-I Thesaurus Bx Classification of Terms

This appendix includes all the terms from the MedSORT-I thesaurus as they were classified according to the MedSORT-I hierarchy, which was built into the MedSORT-II Bx hierarchy. In the file, each term preceded by "***" and followed by ">>". The categorization of the term follows the >>. Note that some terms have more than one category, leading to tangling in the hierarchy. All terms found in this appendix have been added to the MedSORT-II Thesaurus. If a term was already entered in the Thesaurus, only its most specific category was added, in addition to the category given under the MedSORT-II hierarchy.

B18. Index of MedSORT-I Thesaurus Terms

This is an indexing by MedSORT-I hierarchy categories of all those terms from the MedSORT-I Thesaurus, sorted alphabetically by category and term. This was created from the Appendix B17. listing, and demonstrates the fine taxonomy of the MedSORT-I categorization, as each category contains relatively few terms. Determining where each term fits in the hierarchy is greatly simplified by consulting the listing of the fuli MedSORT-II Bx hierarchy, augmented by the MedSCRT-I hierarchy.

C Studies of Semantic Categorization

Cl. Findings: Lexicon, Sorted Alphabetically

This appendix contains a 3,100 line alphabetically sorted listing of the single words found in the INTERNIST-l/QMR findings.

C2. Findings: Lexicon, Sorted by Decreasing Frequency

This appendix contains the single-word types found in INTERNIST-l/QMR findings, in decreasing frequency order.

C3. Findings of Increasing Length

This appendix contains the 4,101 INTERNIST-l/QMR findings sorted into increasing length (number of words per finding).

C4. Phrasal Lexicon of Internist-I/QMR Findings

This appendix lists, in decreasing frequency order, the 2-word and 3-word phrases found in INTERNIST-l/QMR findings. Phrases that occur less than five times are omitted.

C5. Comparison of Human- and Algorithmically-Generated Multi-Words

This appendix lists 1,525 multiwords generated by either human judges or by algorithm (on the basis of sequential probabilities), and shows for each whether it was generated by algorithm or by human or both, and, if generated by algorithm, what its value was.

C6. Grammmar of Quantitative Phrases

This appendix first lists a small grammar (written in SNOBOL) for the quantitative phrases found in INTERNIST-l/QMR findings; and then lists the results of applying it to findings in which such phrases occur.

The SNOBOL specification of the patterns is straightforward: the "I" symbol signifies alternation, and alternation of a pattern with NULL is equivalent to making that pattern optional, as NULL always matches successfully. There is only one recursive pattern, for PER-MEASURE, required because strings like *per centiliter per gram per day per...* can be extended indefinitely.

C7. Grammar of *Pain* Findings

This appendix contains a small grammar for findings containing the word *pain,* written in SNOBOL. Examples of the constituents it detects in such findings are included.

C8. Ambiguously Categorized Terms

This appendix contains information about the numbers of terms assigned two or more categories in the current categorization, and includes: (i) the number of terms assigned to each category; (ii) the number of terms having more than one category; (iii) those terms that are assigned to more than one category; and (iv) the number of terms common to each pair of categories.

C9. Classified Findings

This appendix lists several examples of each of the more occurrent types of classified findings in an easily readable format (those patterns of classified findings that occur at least five times).

This is not the full listing of the more than 5,500 categorized INTERNIST-I/QMR findings. Two forms of the full set of categorized findings are included in the set of files submitted with the Project's report. The first form consists of one classified finding per line, with the sequence of categories of the words in the finding followed by "@@@" and the words themselves. The second form is a readable listing similar to this appendix.

C10. Category Phrasal Lexicon

This appendix lists in order of decreasing frequency the 2- and 3-category sequences found in the categories assigned to words and phrases in the INTERNIST-I/QMR findings. Sequences whose frequencies are less than 5 have been omitted.

C11. Cluster Analysis of Categories

This appendix contains a description of a cluster analysis done on categories, based on their patterns of co-occurrence in INTERNIST-I/QMR findings. Previous analysis was done on a subset of the findings and used an earlier classification scheme; this appendix will give the results of an analysis done with the final classification and the full set of findings.

C12. Trial Grammar of a Subset of Findings

This appendix contains the listing of a trial grammar, written in SNOBOL, to group together contiguous finding words belonging to hypothesized constituent groups, such as "*body-chemical* in *body-fluid*" phrases. The listing is followed by examples of the program's output. The patterns applied differ from the spirit of later constituent definitions most notably in that they incorporate certain corporeal locations into *pathological-factor* phrases, so that, *e.g.*, NOSE MUCOSAL TELANGIECTASIA groups *mucosal* with *teangiectasia* rather than with its associated *body-part*, *nose.*

D Spawn-Related Documents

Dl. SPAWN Source Code

This appendix contains a listing of the source code for SPAWN. It is written in FRANZLISP, but can be easily translated to other LISP dialects.

D2. SPAWN Functions

This appendix contains a detailed description of SPAWN commands. SPAWN commands can be given in two ways: through SPAWN-loop, a user-interface which provides a relatively simple command syntax, although it only allows for a subset of SPAWN functions; and by issuing SPAWN commands directly. The former technique is likely to be used by knowledge-base designers in creating and debugging their data. The latter technique is more useful for controlling SPAWN from a program.

D3. Sample Knowledge Base

This appendix contains a listing of a small knowledge base, selected from our MedSORT-II Project Thesaurus, as implented in SPAWN. It concentrates on the section of the hierarchy built around *substance-techniques.*

D4. An Extended Example of SPAWN-Loop

This appendix contains a listing showing the creation, modification and querying of a small knowledge base, using SPAWN-loop. Each important SPAWN-loop concept is highlighted and explained, when it appears for the first time.

D5. SPAWN User's Manual

This appendix contains a brief user's manual for SPAWN.

E Lexicon and Morphology-Related Documents

El. The Project Lexicon

This appendix contains the Project Lexicon. The Lexicon contains 11402 entries.

E2. Description of Features in the Lexicon

This appendix contains a description of the features (in *features-&nd- values* lists) that appear in the project lexicon, as well as the features generated through MORPH.

E3. MORPH—COMMON LISP Version

This appendix contains the LISP code of the COMMON LISP version of MORPH. The program requires a lexicon and index to run.

E4. MORPH—FRANZLISP Version

This appendix contains the LISP code of the FRANZLISP version of MORPH. The program requires a lexicon to run.

E5. Filter

This appendix contains the LISP code of the filter, as written in FRANZLISP. The filter is to be used together with MORPH and the parser. It serves as an interface between the parser and the user.

E6. Index Generator

This LISP program generates an index from any lexicon in Project-Lexicon form. The index is used by H-P COMMON LISP versions of MORPH. The program contains one function which takes the file name of a lexicon and outputs the index of the lexion as file labeled with the extenstion *.index.*

F Experiments in Natural-Language Processing

F1. FRANZLISP Code for RULEKIT

This file contains FRANZLISP code for RULEKIT, written in 1985 by Jaime Carbonell of the Carnegie Mellon University Computer Science Department for the MedSORT-I Project.

F2. FRANZLISP Code for RULEPAR-II

This file contains FRANZLISP code for RULEPAR-II, a lexically-driven case-frame parser. RULEPAR-II was adapted from RULEPAR, written in 1985 by Jaime Carbonell of the Carnegie Mellon University Computer Science Department for the MedSORT-I Project. RULEPAR-II is the code as adapted for the MedSORT-II Project.

F3. Corpus Parsed by RULEPAR-II

This file contains a listing of the corpus of medical *laboratory-source* observations parsed by RULEPAR-II in the MedSORT-II Project demonstration program.

F4. Sample Parses

This file contains parses as generated by RULEPAR-II when run on the test corpus of medical *laboratory-source* observations.

F5. Frames Used in Demonstration Program

This file contains Mx and Cx frames in SPAWN-readable form that are used by the demonstration system. The frames cover medical *laboratory-source* observations.