Research Paper ■

A Schema for Representing Medical Language Applied to Clinical Radiology

CAROL FRIEDMAN, PHD, JAMES J. CIMINO, MD, STEPHEN B. JOHNSON, PHD

Abstract **Objective:** Develop a representational schema for clinical concepts and apply it to the task of encoding radiology reports of the chest.

Design: The schema was developed following a manual analysis of sample reports from the domain. The schema has two main components: the Medical Entities Dictionary (MED), which specifies the formal representation of the concepts in the domain and of their structures, and the natural-language processor, which specifies the linguistic expressions of the concepts. The schema was evaluated by applying it to a test set of 7,500 reports. Two-hundred reports from the test set were manually analyzed by a medical expert to determine the accuracy and success rate of the system.

Results: 82% of the 7,500 reports that contained relevant clinical information were successfully structured automatically. For the smaller set of 200 reports, 80% were structured successfully with an accuracy rate of 97%.

Conclusions: The schema is a formal representation for clinical concepts in radiology reports, and provides domain coverage that is particularly well-suited for natural-language processing of radiology for use in a decision support system.

■ **J Am Med Informatics Assoc.** 1994;1:233–248.

Automated encoding of the information content of text documents through natural-language processing has long been an important research topic in computer science. The task consists of identifying the occurrence of target terms in the source text given a predetermined target-coded vocabulary. Two substantial obstacles limit the application of natural-language processing to the domain of textual reports of medical procedures. First, no controlled vocabulary exists that can represent all of the variations of medical terms that appear in natural language. Instead of having a single code for "infected pelvic kidney," a combination of codes could be assigned (taking care to differentiate this from "infected kidney pelvis"). Second, simply representing a term in text is inadequate for meaningful encoding, since a determination is also needed as to whether the test is mentioning the term or discussing it; if the latter, the situation in which it is being discussed must be determined (e.g., does the patient have an infected pelvic kidney, a history of it, a family history of it, or has it been ruled out?).

One approach to overcoming these obstacles involves the use of a set of patterns, called schema, by which the processor can recognize the contexts in which terms are appearing. Schemas are needed at the level of understanding individual phrases (infected pelvic kidney vs infected kidney pelvis) as well as at the level of sentence, paragraph, and document structure (differentiating *suspected* diagnoses from *confirmed* diagnoses and/or *ruled-out* diagnoses). Although it is desirable that the methodology involved in developing a schema and its representation be general, particular schemas are domain-specific. They may be developed empirically through analysis of examples of text from the domain or conceptually through

Affiliations of the authors: Queens College of the City of New York (CF) and Columbia University (CF, JJC, SBJ), New York, NY.

Supported in part by grant R29 LM05397 from the National Library of Medicine, by grant 6-61383 from the Research Foundation of CUNY, by a Unified Medical Language System grant from the National Library of Medicine, and by a development contract with IBM Corporation.

Correspondence and reprints: Carol Friedman, PhD, The Center for Medical Informatics, Columbia-Presbyterian Medical Center, 161 Fort Washington Avenue, New York, NY 10032.

Received for publication: 11/12/93; accepted for publication: 1/14/94.

knowledge of the domain. This is an inexact process, at best, that must strike a practical balance between the expressiveness of the language and computational tractability and pragmatics.[1]

At the Columbia-Presbyterian Medical Center (CPMC), we are seeking to encode the information content of the medical text that is included in the clinical-information system to the degree needed for reliable use in our automated decision support system.[2] A controlled vocabulary has been developed[3] as a resource at CPMC, and information must be represented in its terms before being processed by the logic modules in the decision support system. We have developed a language processor that uses the controlled vocabulary as its target and have expanded the controlled vocabulary to include the content of chest-radiology reports.[4] We have also developed a specific schema for use in processing these reports.[5] This paper describes the methodology used for schema development, provides additional details of the schema, and shows the results of applying the schema to a large body of reports.

Background

Formal representations of medical concepts have been proposed by various researchers. Masarie et al.[6] used the notion of concept frames for patient findings to serve as an *interlingua* for unifying disparate controlled clinical vocabularies. The focus of this work was to provide a way to facilitate the translation of one controlled vocabulary into another. The idea was to map the source vocabulary into the interlingua and then to map the interlingua into the target vocabulary. Generic finding frames were used to incorporate potential ways in which a concept may be expressed and also qualified by linguistic modifiers. The authors found that the methodology was promising, but that the construction of generic frames was labor intensive.

A different approach to the representation of medical concepts was adopted by the PEN & PAD prototype.[7] This work is based on the development of a clear understanding of the content of the medical record, with the assumption that it is essential that all clinical information be represented in structured form. The important notion in this approach is to be faithful to the observations of the clinician. The representation includes not only the observations, but also the speculations and suggestions associated with the observations.

The MedSORT project[8] adopted a more linguistic approach to the modeling of medical concepts. Three different levels of information are represented: a linguistic level, a conceptual level, and a contextual level. Lexical items are semantically classified and linked to the conceptual level. Concepts are specified in the form of semantic frames consisting of particular relations among semantic types. In this model, implicit relations are made explicit, and concepts are constrained so that it is possible to specify only well-formed concepts.

Another approach associated with the representation of medical terminologies and coding systems has been proposed by Rossi-Mori[9] in the development of CEN. In this schema, the semantics of medicine is precisely established by specifying well-defined semantic conventions. The focus in this schema is on the organization of terminologic knowledge in medicine in an attempt to establish circumscribed conventions about medical concepts in a given domain.

Our work differs from the other approaches in that it incorporates a natural-language processor. Our schema represents the conceptual levels of clinical information but also includes the linguistic expression of the information along with the necessary linguistic knowledge to encode the linguistic expressions in order to make them consistent with the conceptual levels of information.

The Medical Entities Dictionary

The conceptual levels of information about medical terms are represented within the framework of the Medical Entities Dictionary (MED).[3] The MED is a knowledge base of medical concepts that specifies semantic classificatory information, delineates well-defined semantic relations among concepts, and specifies additional knowledge that is helpful for representing the underlying clinical concepts and for supporting the maintenance and utilization of the controlled vocabulary. The MED forms the heart of the medical representation in the Clinical Information System of CPMC. Clinical applications retrieve patient data using MED concepts. The MED models unique medical concepts, which form a taxonomy of medical classes that supports multiple inheritance. The MED also models structural groupings of information, because these can also be considered clinical concepts. In general, concepts that are higher in the hierarchy correspond to abstract entities that specify the overall structure of the information, such as **Laboratory Test** and **Diagnosis**, and therefore many of these concepts are not seen in the actual textual reports. The concepts that are lower in the hierarchy are generally associated with more familiar clinical concepts, such as **serum glucose test** and **diabetes**

mellitus, which commonly occur in the textual reports. Currently, the MED contains over 34,000 concepts.

The MED attempts to represent both the medical concepts as well as the structure of the information in the concepts. Therefore, if the information in the clinical radiology reports could be entered in structured form using only controlled-vocabulary concepts, no further knowledge would be needed. However, in order to perform natural-language processing of the text (i.e., the mapping of the text to unique concepts), a lexicon and grammar, along with other knowledge components, are required.

The Natural-language Processor

The natural-language processor translates the text into well-defined MED concepts. In order to accomplish the task, the processor utilizes three separate functional phases—the parsing phase, the compositional phase, and the encoding phase—that employ different linguistic sources of knowledge. The parsing phase generates a preliminary structured form from the source text, the compositional phase regularizes the preliminary structures further, and the encoding phase maps the regularized structures into the controlled-vocabulary concepts that are maintained in the MED. The components are summarized below, but a more detailed description is available.[4]

The Parsing Phase

The first component, the parser, transforms the text into a preliminary structured form where the text terms have been translated into standardized target forms. A semantic grammar and lexicon are used for this task.

The lexicon semantically categorizes single words and multi-word phrases in the domain and specifies their target forms. The target forms are the standard output forms associated with the words or phrases. For example, the word *lung* is a body location type word and its target form is **lung**, which is a MED concept that happens to have the same name. The word *mediastinal* is also a body location type word but its target form is **mediastinum**.

The lexicon uses a small subset of the classes defined in the MED. These classes tend to be fairly general rather than finely detailed because they are based on distributional patterns observed in actual reports. This simplifies the task of creating lexical entries and avoids overlapping classes and inconsistencies in classification. For example, in the lexicon, there is only one coarse class, **Bodyloc**, which represents organs, body regions, and areas. However, in the MED, the hierarchical organization of the concepts is finer. For example, **Coronary Artery** is classified in the MED as a **Blood Vessel** that is a subclass of **Bodyloc**. In this way, the more complex and dynamic hierarchical organization of semantic classes according to concepts in the MED will not have an impact on the lexicon.

The grammar models the semantic relations found in the domain. It determines the structure of the text by specifying well-formed semantic co-occurrence patterns of the domain along with their corresponding target structures, which are compatible with the structure of the findings as modeled by the MED. For example, the simplified co-occurrence pattern

DEGREE CHANGE PATHOLOGIC_ENTITY

represents findings consisting of a common pattern that contain a sequence of words or phrases corresponding to the three semantic classes. For example, in *slight increase in congestion, slight* is a degree type word, *increase in* is a phrase denoting change type of information, and *congestion* corresponds to a pathologic condition. This pattern is interpreted by the grammar so that the degree information *slight* qualifies the change information *increase in*, which together qualify the pathologic condition *congestion*, which is the finding. The appropriate target structure for this pattern is specified in the grammar along with the pattern, enabling the parser to transform the well-formed patterns appropriately. For example, the target structure for the above pattern will be specified in the grammar so that there is a finding that corresponds to **PATHOLOGIC_ENTITY** (*congestion*) and that is qualified by a change qualifier corresponding to **CHANGE** (*increase in*). Similarly, the target change qualifier will have a degree qualifier corresponding to **DEGREE** (*slight*).

Another common semantic pattern consists of the sequence **DESCRIPTOR BODYLOC PATHOLOGIC_ENTITY** as in *nodular right upper lobe opacity*. The target form for this sequence is specified so that the pathologic entity *opacity* is a finding that has two qualifiers, a body location qualifier *right upper lobe* and a descriptive qualifier *nodular*.

The Compositional Phase

The second phase of processing is called the compositional phase because it further regularizes the structured output by identifying and composing multiword phrases that were separated in the original text. For example, the words in the phrase *nodular opacity* may appear in a report as *nodular right upper lobe opacity*.

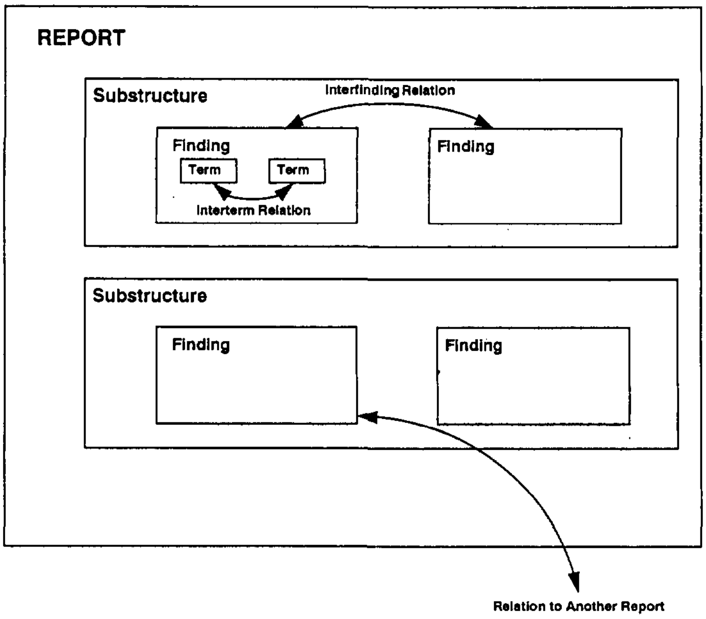


Figure 1 An overview of the concepts and interrelationships associated with the structure of a report. A report consists of **substructures** that correspond to different sections of the report. The **substructures** contain **findings**. A **finding** may be related to another **finding** in the same report, as represented by the relation **Interfinding Relation**, or it may be related to another report, as represented by the arrow pointing to another report. Findings contain **terms** that are also related.

This phase utilizes mappings that model the compositional structure of multi-word phrases and specify their corresponding standard forms. The compositional mapping of a multi-word phrase is generated by the natural-language processor when the phrase is added to the lexicon. The processor treats a multi-word phrase just like a sentence in a report and structures it in the usual manner; however, after the phrase is processed, its structured form is saved. A mapping for the phrase *nodular opacity* consists of a structure containing a finding *opacity* with a descriptive qualifier *nodular*. The preliminary output produced by the preceding parsing phase for *nodular right upper lobe opacity* will be changed in this phase so that the finding is associated with *nodular opacity* instead of *opacity*.

The Encoding Phase

The third phase of processing maps the regularized structured output to encoded form. The ability to compose multi-word phrases significantly reduces the variety of expression that is encountered in text. However, frequently there are synonymous phrases that have different target forms. The synonym knowledge base provides the knowledge that is needed to associate the regularized terms with concepts in the MED, and therefore it links the language of the text to the well-defined concepts in the domain. For example, the synonym knowledge base associates *nodular opacity* with the MED concept **nodule**.

The synonym knowledge base is also used to specify domain-specific default information for particular domains. For example, in the domain of chest x-rays, the term **right upper lobe** refers to **right upper lobe of lung**. In this way, body location information that is frequently missing (because it is implied by the type of examination) can be explicitly supplied.

After this phase is completed, the structured form of *nodular right upper lobe opacity* will be completely encoded so that the finding is associated with the concept **nodule** and the body location qualifier is associated with the concept **right upper lobe of lung**.

Methods

Manual Analysis of Sample Reports

A manual analysis of sample radiology reports (the training set) that were randomly retrieved from the on-line textual radiology database at CPMC was performed to detect and represent medical concepts. The analysis was performed with the view that in order for codified radiology reports to be useful for applications such as decision support, it is necessary to represent the context of findings as well as the findings themselves. The schema generally attempts to represent context through the use of additional relations. For example, even the most superficial examination of radiology reports reveals a structure with components such as **Clinical Information and Impressions**, which explicitly represent the section of the examination that the clinical information was found in. The rationale for this approach is that it may be desirable to retrieve concepts when they occur in multiple contexts and then filter out the undesirable contexts, if any. The direct implication of this approach is that concepts are needed that represent the structure of the reports in addition to the contents. Figure 1, which is described below, is an overview of the concepts and the interrelationships that are associated with the structure of a report.

The analysis of the radiology reports was performed in several different steps:

1. The first step in the analysis was concerned with determining the overall structure of the reports. This involved the high-level analysis of the report itself as well as the analysis of the complex interrelationships among the concepts found in the report. This

level of information consists of identifier information, such as the time and type of the examination, and also determines the context of the findings in the report. For example, there are different sections in the report structure, such as **Clinical Information**, **Description**, and **Impression**, which affect the way the information will be used. This level is represented in Figure 1 as the substructures that constitute the report.

2. The second step consisted of an analysis of the interrelationships among individual findings. Findings are often related to other findings, as in *hazy opacities over both lung fields may represent pleural effusion or interstitial edema*. In Figure 1, this type of relationship is represented by the inter-finding relation that is shown connecting two different findings in the same report, and by the relation that connects a finding in one report to that in another report.

3. The third step was an analysis of the structure of individual findings in the reports. This involved determining the underlying broad semantic categories and then the semantic relationships used to express the findings. The semantic categories were determined based on Systematized Nomenclature of Medicine (SNOMED) axes[10] and on work of the Linguistic String Project.[11] The semantic relations were determined based on analyzing the reports to find common semantic co-occurrence patterns among the semantic categories.

This level of information consists of complex structures of interrelated medical concepts. Findings generally consist of central concepts with modifiers. For example, in *large opacity in base of left lower lobe, opacity* is the central concept of the finding, and *large* and *base of left lower lobe* are size and body location qualifiers. In Figure 1, the findings are contained in the substructures of the report, and the relation between terms in a finding is represented as the interterm relation.

4. The fourth step in the analysis involved specifying the remaining concepts in the domain along with the hierarchical orderings. The concepts in the domain were determined based on semi-automated techniques. For this step, a large collection of the impression section of radiology reports (2 mb of text) was used to obtain the frequency of words and to suggest multi-word phrases based on statistical techniques. The hierarchical ordering was performed by a medical expert and was based on expert knowledge of the domain.

This level of clinical information consists of the well-defined medical concepts that make up the findings. These concepts have a unique meaning, although in language they are generally expressed in a variety of ways. For example, *cardiomegaly, enlarged heart, cardiac enlargement,* and *enlarged cardiac silhouette* represent the same concept for our application within the context of radiology findings.

5. The fifth step in the analysis involved the development of the components for the natural-language system and then the application of natural-language processing to the sample reports to extract salient information and to map it to a form consistent with the schema. This resulted in several rounds of refinements.

This fifth level of information is the linguistic level. This consists of lexical information associated with the words and phrases used in the reports to express the underlying medical concepts, and it also consists of enumerable well-formed co-occurrence relations among the words and phrases that express the semantic relations among the concepts. The words and relations may be ambiguous and therefore do not necessarily correspond to unique, well-defined concepts. For example, the word *radiation* could mean **radiation therapy** or **x-ray examination**, depending on the context. Similarly, the semantic relation expressed by *upper and lower right* could be interpreted so that *upper* and *lower* both qualify *right* or so that only *lower* qualifies *right*.

Developing Components Used by the Natural-language Processor

Several components were developed subsequent to the manual analysis, as outlined below:

- Formal representations of the structures of both the reports and the findings were developed and represented within the framework of the MED. These are discussed in detail in the first two sections of Results.
- A controlled vocabulary of well-defined medical concepts associated with radiology reports of the chest was developed and added to the MED, and a hierarchical ordering of the concepts was specified. The vocabulary is discussed in Results.
- A formal grammar and lexicon for natural-language processing was developed. The grammar reflects the information patterns underlying the structure of the findings in the reports by specifying well-formed semantic co-occurrence relations among the clinical terms in the reports in addition to translating them and mapping them within the target structure, which is the formal schema.

A lexicon was created for the words and phrases in the reports consisting of semantic classificatory information and corresponding target forms compatible with the MED.

- A compositional component was created automatically by processing the decomposable multi-word phrases using the natural-language processor to produce the structured representation of the phrases. This was possible because these phrases resemble findings in the radiology text. Since the processor can structure the text, it can also structure the multi-word phrases.
- A knowledge base of synonyms was created associating standard target forms with MED concepts.

Evaluation

The schema was evaluated using a test set of 7,500 reports, containing a total of 13,767 sentences. The coded representation of the findings from the test set of actual reports was automatically produced by the language processor. The majority of the reports were obtained from the textual radiology database maintained at CPMC, but several reports obtained from other sites were also added to the test set.*

A random sample of 200 reports (containing a total of 371 sentences) from the test set of 7,500 reports was selected and the corresponding output forms were evaluated manually by a radiologist in the Radiology Department of CPMC to determine whether they correctly modeled the clinical information in the reports. Prior to the evaluation, the radiologist was trained on the same training set that the natural-language processor was trained on. As part of the manual evaluation of the 200 reports, some sentences were determined to be irrelevant to this study because they contained notes and comments, such as *Dr. Smith was notified* and *see above*. Although these sentences may contain some clinical information, they were ignored in this study because they did not contain the type of information used for decision support.

Two measures used to evaluate the performance of the system were encoding success rate and accuracy. In this study, encoding success rate is defined as the number of sentences for which structured output forms were generated divided by the total number of sentences that contained clinical information. Accuracy is defined as the number of sentences containing clinical information that were correctly structured by the system divided by the total number of sentences that were structured.

Representation

We use the formalism of conceptual graphs (CGs)[12] to represent the concepts. This choice was made based on the readability and straightforward semantics of CGs, and also because of the increasing popularity of CGs within both the medical-informatics community and the database and knowledge-base communities.

In the CG formalism, a concept is represented as a set of specified relations among other concepts in the MED. A concept is enclosed in square brackets and followed by the relations associated with it. Each relation appears in parentheses and is followed by an arrow (→). The relations are indented for readability. The values that each relation can take are specified by another concept that appears in square brackets after the arrow. Thus, the general format of a concept with N relations is:

```
[Concept]-
  (Relation1)->[Concept1]
  (Relation2)->[Concept2]
  :
  (RelationN)->[ConceptN].
```

Notice that the main concept is followed by a dash (-), and that the definition is terminated by a period (.). The number of values that a relation is permitted to have (its cardinality) is indicated by including a constraint following the related concept's name (the concept occurring after the arrow). In rough terms, the constraint :{*} means that the relation must have 0 or more values; :{*}**@>0** means that the relation must have 1 or more values; and :{*}**@<2** means that the relation must have 0 or 1 values. When no constraint is specified, the default cardinality is assumed to be exactly 1.

Results

The Report Structure

Analysis of the radiology reports revealed that the reports are more than a collection of concepts. They have a structure that places the concepts in a variety of contexts. In these contexts, the concepts retain their meanings but their implications for uses such as decision support may vary.

Contexts for radiology findings include: information about the patient given to the radiologist, observa-

*These reports were chosen by several other groups also involved in developing schemas for representing clinical information, and are useful for the purpose of facilitating comparison and evaluation of the various schemas.

tions made in the description of the film, and interpretations based on the description. The same findings can appear in any of these contexts (often in the same report) and therefore a single schema is used to represent them. Concepts may also appear in radiology reports in two other contexts: as clinical information about the patient (current or past conditions) and as reasons for the examination (such as conditions the ordering physician wishes to rule out).

Reports usually have several fields dealing with information about the patient, the person who ordered the test that produced the report, the person who dictated the report, and the person who transcribed the report, as well as dates for the procedure and report. They are straightforward to represent and are not included here. Thus, for the purposes of this paper, reports have simplified descriptions, which are shown in Figure 2. This figure is a description of the structure of the information in the overall report. The structure of a report is represented in the canonical-graph **Report**. It consists of two relations, **Procedure Type** and **Identifier Info**, which link to the concepts called **Procedure** and **Event Identifier**, respectively. A canonical graph, **Chest Xray Report**, is instantiated when an actual report is represented. The next two CG definitions in Figure 2 show successive refinements of the type **Report**. **Xray Report** and **Chest Xray Report** are both descendants of **Report**, and therefore they automatically inherit the relations **Procedure Type** and **Identifier Info**. **Xray Report** refines the relation **Procedure Type** because it is associated with the more specific class **Xray Procedure** instead of with the class **Procedure**. Most of the relations introduced in **Xray Report** are unmodified when inherited by **Chest Xray Report** and are therefore not shown. The exception is the relation **Procedure Location**, whose related concept is refined to **Chest** in **Chest Xray Report**.

Two of the relations in **Xray Report** allow for the inclusion of clinical information supplied by the person who orders the x-ray procedure: **Clinical Information** (such as patient age, gender, symptoms, and known conditions) and **Reasons for Exam** (such as change in a finding or ruling out a disease). The two reports in the sample data do not include this information so they are not modeled further. The **Comparison Report** relation allows for the inclusion of a reference to a previous report, if such appears in the current report. The final two relations, **Description** and **Impression**, correspond to the two sections of radiology reports that typically contain sentences consisting of radiology findings. The **Radiology Finding Sentence** concept represents the original sentence in the Description or Impression section of a radiology report in addition to its structured form. The relation **Text** has as its domain **String Data**, which is a primitive concept used to represent the actual textual sentence(s) of the report. **Structured Finding** relates to the structured form of the relevant clinical information in the sentences, which is called **Rad Finding**. **Rad Finding** consists of the complex arrangements of medical concepts and corresponds to individual findings in the text with modifiers.

```
[Report]-
    (Procedure Type)->[Procedure]
    (Identifier Info)->[Event Identifiers].

[Xray Report] -
    (Procedure Type)->[Xray Procedure]
    (Procedure Location)->[Bodyloc]
    (Clinical Information)->[Finding:{*}]
    (Reason for Exam)->[Motivations:{*}]
    (Comparison Report)->[Xray Report:{*}]
    (Description)->[Radiology Finding Sentence:{*}]
    (Impression)->[Radiology Finding Sentence:{*}].

[Chest Xray Report]-
    (Procedure Location)->[Chest].

[Radiology Finding Sentence]-
    (Text)->[String Data]
    (Structured Finding)->[Rad Finding:{*}].
```

Figure 2 The concepts **Report**, **Xray Report**, **Chest Xray Report**, and **Radiology Finding Sentence** are high-level concepts that specify the overall structure of a report. These concepts contain relations such as **Procedure Type**, **Procedure Location**, and **Clinical Information** that provide the context for the findings in the report.

The Structure of the Findings

Findings

An analysis of radiology findings showed them to be complex arrangements of basic medical concepts. The possible permutations of radiology findings suggest that enumerating them would be impractical. However, the interactions of concepts in each radiology finding appear to be of a relatively small number and are represented by one concept **Rad Finding**, which is shown in Figure 3. A **Rad Finding** has a central finding represented by the relation **Central Finding** that is qualified by modifiers represented by the relation **Bodyloc Mod** and **Finding Mod**, which are shown in Figures 4 and 5. Thus, Figures 3, 4, and 5 represent the components of the overall report schema. The other relation in **Rad Finding** that may occasionally occur in a report is called **Evidential Procedure**. This relation represents evidence of sur-

```
[Rad Finding]-
    (Central Finding)->[Rad Finding:{*}]
    (Bodyloc Mod)->[Bodyloc:{*}]
    (Finding Mod)->[Modifier:{*}]
    (Related Finding)->[Relational Finding:{*}]
    (Evidential Procedure)->[Surgery:{*}]
    (Technique Information)->[Technique:{*}]
    (Management Information)->[Management Procedure:{*}].

[Relational Finding]-
    (Relation)->[Interfinding Relation:@1]
    (Structured Finding)->[Rad Finding:@>0].

[Interfinding Relation]-
   (Relation Concept)->[Interfinding Relation:@<2]
   (Certainty Mod)->[Certainty:@<2].
```

Figure 3 The **Rad Finding** concept models the structure of the findings in the report. Generally the findings consist of the central finding relation **Central Finding** with qualifiers **Bodyloc Mod** and **Finding Mod**. The relation **Related Finding** represents secondary related findings that are usually implied.

gical procedures, such as **mastectomy** and **lobectomy**, that are stated in the reports. Another relation, **Technique Information**, represents information in the reports that concerns technical issues dealing with the x-ray itself and therefore is not actually a finding. For example, **expiratory film** and **poor inspiration** are associated with technical issues and do not correspond to findings. This relation is represented in a radiology finding because it is often stated in conjunction with other findings, as in *infiltrate cannot be excluded because of poor inspiratory effort*. Another relation called **Management Information** represents patient-management information, such as **follow-up suggested** and **clinical correlation suggested**, which occasionally appears on the report.

Radiology findings interrelate in very limited ways. For example, a radiology finding may reference another radiology finding mentioned in a previous radiology report. This relationship is represented by the **Finding Mod** relation that has a temporal modifier as its value. If the finding was previously mentioned, as in *markings are noted again*, the value of the temporal modifier would be **again**. For a comparison between findings, as in *opacity has increased*, the value of the temporal modifier would be **increase**.

Another way in which radiology findings may be related is when one radiology finding suggests a second radiology finding, as in *markings are consistent with atelectasis*. Since the second radiology finding is not a direct observation in the report, but rather is included as related to the first radiology finding, it is included as part of the description of the first radiology finding rather than as a parallel finding in the report. This variety of related finding is represented by the relation **Relational Finding** in **Rad Finding**. **Relational Finding** has its own structure, which is also shown in Figure 3. **Relation** represents the underlying relationship that connects the interrelated findings, such as **compatible with**, **may represent**, and **may be related to**. The relation **Structured Finding** represents a nested or secondary finding. This has the same structure as the primary finding. The concept **Interfinding Relation** represents relations that occur between findings, and it may be

```
[Bodyloc]-
    (Primary Loc)->[Bodyloc:{*}]
    (Spatial Mod)->[Spatial Relation:{*}]
    (Bodyloc Mod)->[Bodyloc:{*}]
    (Region Mod)->[Region:{*}]
    (Orientation Mod)->[Orientation:{*}]
    (Quantity Mod)->[Quantifier:{*}].
```

Figure 4 The **Bodyloc** concept models the structure of body location qualifiers.

```
[Modifier]->
    (Change Mod)->[Change:@<2]
    (Degree Mod)->[Degree:@<2]
    (Certainty Mod)->[Certainty:@<2]
    (Status Mod)->[Status:@<2]
    (Descriptor Mod)->[Descriptor:*].

[Change]-
    (Change Concept)->[Change:@<2]
    (Certainty Mod)->[Certainty:@<2]
    (Degree Mod)->[Degree:@<2]

[Certainty]-
    (Certainty Concept)->[Certainty:@<2]
    (Degree Mod)->[Degree:@<2].

[Degree]-
    (Degree Concept)->[Degree:@<2]
    (Degree Mod)->[Degree:@<2].

[Status]-
    (Status Concept)->[Status:@<2]
    (Certainty Mod)->[Certainty:@<2]
    (Procedure Mod)->[Procedure:@<2].

[Descriptor]-
    (Descriptive Concept)->[Descriptor:{*}]
    (Descriptor Mod)->[Modifier:{*}].
```

Figure 5 These are finding qualifiers that are not body locations. These qualifiers model change, certainty, degree, status, and descriptive types of information.

modified by a certainty modifier. For example, the relation *possibly related to* has a central relation concept **related to** and a certainty modifier *possibly*. Findings that are parallel, as in *markings and opacity noted, markings as well as opacity noted*, and *markings with opacity*, are represented simply as multiple findings on the same level.

Body Location Qualifiers

The **Bodyloc Mod** relation, which is illustrated in Figure 4, represents the body location information in the radiology reports. The **Bodyloc** structure consists of a relation **Primary Loc** representing the primary body location. Thus, if the sentence is *heart is enlarged*, the primary body location relation would be the concept **heart**. The primary location can also have modifiers, which are also shown in Figure 4. The relation **Spatial Mod** represents the prepositional or adverbial relation associated with the **Primary Loc** slot. If the radiology finding is *opacity in left lung*, the spatial relation is **in**; if the finding is *opacity under base of left lung*, the spatial relation is **under**. To simplify retrieval from this structured form, **Spatial Relation** may be missing, in which case it may be assumed that its value is a member of the type **contiguous** that corresponds to the terms *in, on, at*, and *along*. The implication is that the finding is located contiguous to the specified body location. If **Spatial Mod** is present, its value changes the meaning of the location of the finding. For example, when the value is **under**, the central finding is not in the specified body location but under it. Therefore, it is used as a relative positional locator in order to specify where the finding actually is relative to a body location.

The **Bodyloc Mod** relation represents a body location modifier of the primary body location. In *finger of right hand*, the primary location would be **finger**, modified by a body location **right hand**. The relation **Region Mod** represents a relative body location modifier of the primary body location. A region is a relative area of a body location, such as *upper, left*, and *base*. The **Orientation Mod** relation represents the orientation of the primary body location. In *transverse heart*, the primary body location is **heart** and the orientation modifier is **transverse**. The **Quantity Mod** relation represents a quantifier, such as **2** in *2 fingers*.

Remaining Qualifiers

The other modifiers of a finding are represented by the relation **Finding Mod**, which is associated with the concept **Modifier**. This consists of information that modifies the radiology finding, such as temporal, certainty, degree, and quantity. For example, the Rad Finding for *severe chronic pleural effusion* would have a Central Finding **pleural effusion**, a severity modifier **severe**, and a status modifier **chronic**. The concept **Modifier** has five subtypes representing degree, change, certainty, status, and descriptive information.

The concept **Change** represents information denoting a change in the finding. It is assumed that the finding still exists, but has changed (or has not changed). Examples of this type of modifier are **improved, worsened, no change**, and **decrease**. **Change** information may be modified by degree and certainty information, as in *slight increase* and *possibly increased*. The concept **Status** represents other types of temporal information, including the resolution of a finding and new findings. Examples of status modifiers are **resolved, new, chronic**, and **previous**. Status information could also represent the finding in relation to a procedure that was performed. This is represented by the relation **Procedure Mod**. Some examples of this type of information are *post mastectomy status, status post coronary artery bypass graft*, and *post operative status*. Status information may also have a certainty modifier, as in *possibly chronic*.

The concept **Certainty** represents certainty modifier information, as in *possible cardiomegaly* and *no evidence of cardiomegaly*. Some terms corresponding to this type of information are: *no evidence of, most likely, probable*, and *unlikely*. Certainty information can have degree modifiers, as in *highly likely*. Although the MED models nested certainty information, the natural-language processor produces target structures that have limited values for certainty information (i.e., **no, low certainty, moderate certainty, high certainty**, and **undetermined**). This approach was chosen because this information is basically imprecise. In addition, restriction to a limited set of values greatly simplifies the writing of automated queries.

The concept **Degree** corresponds to degree of severity information, as in *severe cardiomegaly*. Terms corresponding to this type of information include *slightly, extensive*, and *moderate*. Degree modifiers may be modified by other degree modifiers, as in *very mild*.

Another modifier concept is called **Descriptor**. It is used to represent modifier information that does not fall into any of the other classes. Examples of terms denoting this type of information are *linear, primary, original, prominent*, and *round*. Finer classification of modifiers may make this type unnecessary.

The Concepts in the MED

Every concept in the MED is a generic concept and as such should be regarded as a type or class. Instantiations of report concepts along with the findings are created when the reports are structured. In our system, this occurs when the structured forms are produced by the natural-language processor. The structured data will eventually be stored in the coded clinical patient database at CPMC[13,14] when the natural-language processor is integrated with the clinical information system (CIS). Thus, the database will contain instantiations of the concepts, whereas the MED contains the representations of the concepts themselves. In order to differentiate between a canonical graph, which corresponds to a generic concept, and an instance of a concept in CG notation, a concept instance is shown with an identifier. For example, the CG for **Chest Xray Report** shown in Figure 2 represents the generic concept **Chest Xray Report**, whereas the CGs for **Chest Xray Report** generated from actual reports will contain identifiers because they represent instances of the concept **Chest Xray Report**.

It is difficult to specify what the appropriate granularity for the concepts in the reports should be since the granularity is dependent on the need for generic classes within particular applications. Thus, we propose a hierarchy of concepts that we think will be helpful for decision support. However, others may view the hierarchical organization differently if the intended application is different. An extension of the hierarchy is expected as more applications use the concepts.

Figures 6 and 7 list some of the concepts in the MED along with their hierarchical orderings. In CG formalism, $A < B$ specifies that A is a subclass of B. Figure 6 shows that **Medical Entity** is the highest concept in the hierarchy; under it are the concepts **Report, Finding, Bodyloc, Therapeutic Device, Procedure,** and **Modifier**. The concept **Modifier** has different types of modifiers as subclasses.

To facilitate decision support applications that use structured findings from natural language, it is convenient to divide certain modifier classes into a limited number of subclasses. For example, the certainty class is divided into five subclasses—high certainty, moderate certainty, low certainty, negation, and undetermined. These classes represent fuzzy concepts that are frequently found in the reports, such as *appears to be, is possibly,* and *may represent,* and are used to hedge the certainty of findings. In this way, applications using the controlled vocabulary do not have to enumerate all the terms belonging to each certainty class, but could retrieve information associated with certainty based on one of the five subclasses. Having classes that represent fuzzy concepts does not conflict with the goal of having well-defined concepts because vague information occurs frequently in the reports and therefore it is important to represent this type of information. These concepts are viewed as being well-defined because their meanings are understood to specifically denote information that frequently occurs in language and that is not precise.

```
Chest Xray Report < Xray Report < Report < Medical Entity
Rad Finding < Finding < Medical Entity
Bodyloc < Medical Entity
Therapeutic Device < Medical Entity
Procedure < Medical Entity
Modifier < Medical Entity
Cfinding < Rad Finding
Descriptor < Modifier
General Descriptor < Modifier
Region < Modifier
Orientation < Modifier
Certainty < Modifier
Low Certainty < Certainty
Moderate Certainty < Certainty
High Certainty < Certainty
Negation < Certainty
Undetermined < Certainty
Degree < Modifier
Low Degree < Degree
Moderate Degree < Degree
High Degree < Degree
Temporal < Modifier
Status < Temporal
Change < Temporal
Xray Procedure < Procedure
Surgery < Procedure
```

Figure 6 Examples of high-level concepts in the hierarchy. These concepts model the structure of the information in the reports and high-level conceptual classes useful for retrieval applications. Therefore, these concepts are not explicitly observed in the reports.

In the formal schema, the meaning of the central finding is always the same but the interpretation of the finding varies depending on the various modifiers associated with the finding. For example, the difference between **High Certainty** or **Moderate Certainty** qualifying a finding may not be significant for most applications, but having a modifier whose value is **No** will definitely make a major difference in whether the finding should be retrieved or not. For example, *no evidence of cardiomegaly* is significantly different from *possible cardiomegaly* and *extensive cardiomegaly* when related to applications retrieving patients who have (or may possibly have) cardiomegaly.

The Degree modifier also has three subclasses—**Low Degree, Moderate Degree,** and **High Degree**—to facilitate retrieval. Having Certainty, Degree, Change, Status, and Bodyloc modifiers inherently complicates the retrieval of the information; however, by subclassifying these types so that they consist of only a few alternative subclasses, retrieval is greatly simplified. The medical concepts shown in Figure 7 are towards the bottom of the hierarchy. In order to save space, they are shown as lists associated with the appropriate classes. The ** in Figure 7 indicates that the concept is decompositional (all of the words are not always contiguous in the reports), and therefore a compositional structure should also be created in order to facilitate natural-language processing.

Notice that hierarchical subclasses in the MED are created as needed. For example, **Blood Vessel** is a subclass of **Bodyloc** and **Coronary Artery** is a subclass of **Blood Vessel**. In this way, applications could retrieve information based on varying degrees of granularity, i.e., ranging from a rather generalized concept, **Blood Vessel**, to a more specific concept, **Coronary Artery**.

Modeling the Compositionality of Concepts

In order to support natural-language processing and the ability to map one controlled vocabulary to another, a compositional modeling of lexically complex concepts is desirable. These concepts are generally multi-word concepts that can appear in a variety of forms where the words of the concepts may be separated from each other. Figure 8 shows how the compositional structures of two concepts are modeled in the MED. For example, **Blunting of Costophrenic Angle** contains a **Central Finding**, which is **Blunting**, and a **Bodyloc Mod**, which is **Costophrenic Angle**. Thus, if this concept is expressed in a report in a variant form such as *costophrenic angle appears slightly blunted*, it will be possible to recognize it by matching the structured form of the variant expression to the structured form of **Blunting of Costophrenic Angle**.

Other concepts, such as **lung**, **atelectasis**, or **circumflex artery**, are not lexically decomposable because they are either single words or multi-word terms that are always together in the text. The structure of each decomposable term was obtained by automatically processing and structuring the term using the natural-language processor. This is possible because multi-word terms are just like typical text sentences that contain findings.

Results of Evaluation

For the set of 200 reports associated with the manual evaluation, the encoding success rate of the natural-

Figure 7 Examples of lower-level concepts in the hierarchy, which generally correspond to clinical terms observed in the reports.

**Blunting of Costophrenic Angle **, Cardiomegaly, Distended Pulmonary Vessels ** < Cfinding
Pleural Effusion **, Pleural Fluid **, Congestive Heart Failure, Atelectasis < Cfinding
Pneumonia, Consolidation in Lung ** < Cfinding
Opacity, Effusion < Pfinding
Blunting, Distended, Enlarged < Descriptor
Clip, Sternotomy Wire, Surgical Clip, Wire Clip < Therapeutic Device
Bypass Surgery, Coronary Artery Bypass Surgery, Lobectomy, Sternotomy < Surgery
Radiation Therapy < Xray Procedure
Chest, Fissure, Heart, Artery Distribution, Lung, Lower Lobe of Lung **, Hilum < Bodyloc
Mediastinum, Blood Vessel, Pleura, Paramediastinum, Costophrenic Angle < Bodyloc
Distribution of Circumflex Artery < Artery Distribution
Pulmonary Vessel **, Coronary Artery, Circumflex Artery < Blood Vessel
Right, Left, Lower, Mid, Base < Region
Anteroposterior, Transverse, Lateral < Orientation
Intact, Platelike < General Descriptor
cannot exclude, less likely < Low Certainty
possibly, maybe < Moderate Certainty
evidence of, probably, most likely, seen, noted, present < High Certainty
no < Negation
cannot evaluate < undetermined
Again, New, Persistent, Chronic, Previous, Postoperative < Status
Decrease, Increase, Improved, Changed, No change < Change
slight, some, minimal, low < Low Degree
moderate < Moderate Degree
severe, extensive, increased < High Degree
multiple, many, few, single < Quantity Object
related to, consistent with, compatible with < Interfinding Relation
in, on, at, along < Contiguous Spatial Relation
behind, below, above, under, near, adjacent to, next to < Non-Contiguous Spatial Relation
poor inspiration, expiratory film, suboptimal study < Technique
followup suggested, clinical correlation, suggested < Patient Management**

```
[Blunting of Costophrenic Angle]-
    (Central Finding)->[Blunting]
    (Bodyloc Mod)->[Costophrenic Angle].

[Left lower lobe of lung]-
    (Primary Loc)->[Lung]
    (Region Mod)->[Left lower lobe].
```

Figure 8 Two examples demonstrating compositional mappings that are associated with the structure of multiword phrases. The concepts **Blunting of costophrenic angle** and **left lower lobe of lung** may occur in the reports in variant forms, as in *blunting of the right lateral costophrenic angle* and *opacities in left and right lower lobes of lung*.

language processor was 80%. Eleven percent of the sentences could not be processed because they contained words or phrases that were not in the lexicon. Two percent of the sentences could not be processed because of spelling errors. Seven percent of the sentences could not be processed due to shortcomings in the grammar. These sentences contained atypical semantic patterns or patterns that were too complex for the semantic grammar to handle.

The accuracy of the processor was 97%. The majority of inaccuracies are considered *minor* errors within our application because it is unlikely that they will affect retrieval of the relevant information for the decision support application; however, these inaccuracies may be more serious within other applications requiring more detailed qualifier information. For example, in *left and right base of lung*, the parser interpreted the qualifier information as if the sentence were *left lung and right base of lung*. A more accurate interpretation is *left base of lung and right base of lung*. Most of the inaccuracies were due to the incorrect structuring of qualifier information by the natural-language processor. For example, in *diminution of pneumothorax but small residual*, the phrase *diminution of pneumothorax* was structured correctly, but *small residual* was structured as if it were an independent finding so that the information *small residual*, qualifying *pneumothorax*, was lost.

The overall encoding success rate for those reports that were automatically processed was also calculated. The encoding success rate of the processor was 82%. Again, undefined words and spelling errors accounted for most of the failures (12%); only 6% of the sentences could not be encoded due to grammar problems. The output was analyzed by one of the authors to categorize the different types of problems encountered by the system. A very small number of problems (less than 1%), which are described in the Discussion section, were due to the simplifications that were made in the representation of the concepts.

A Sample Report

Figure 9 shows an instance of a structured form for one of the reports that was used by other groups also involved in developing representation schemas for clinical information. For brevity, the only identifiers shown are associated with the concepts **Radiology Finding Sentence** and **Rad Finding Structure**, and they correspond to the sentence identifier **#bw22.4.1** and the finding identifiers **#bw22.4.1.1** and **#bw22.4.1.2**.

In Figure 9, the actual text sentence is shown as the value of the relation **Text**, and the structured findings are shown under the relation **Structured Finding**. Each finding is assigned a unique identifier. The **Central Finding** relation of the primary finding consists of the concept **Opacity** that has a **Bodyloc Mod** relation whose value is a **Bodyloc** where the primary location is **Mediastinum. Mediastinum** has a **Region Mod** that is **right** and a **Spatial Relation** modifier whose value is **near**, signifying that the **opacity** is not located on the **mediastinum** but near it. There are two **Finding Mods** associated with **opacity**. One contains the status type information **persistent**, signifying that the finding is not new, and the other contains the degree type of information **increased**, meaning **above normal degree**. The particular type of **Finding Mod** is not shown because the concept **increased** is well-defined in the MED, and therefore its semantic class is known. The structured finding has a secondary or related finding **Relational Finding** whose value is the relation **related to** with a certainty modifier **Moderate Certainty**. If convenient for applications, relational concepts such as **related to, suggestive of, consistent with**, and **compatible with** may be put into one concept superclass. The secondary finding is **radiation** (in the sense of radiation therapy), and a modifier **previous**. The remaining findings of the sentences for sample reports have been processed and are shown in Appendix A.

Discussion

The evaluation study was designed primarily to evaluate the natural-language processor. However, the processor is closely related to the conceptual model. When the processor automatically transforms the text into a structured output form that is determined by a clinical expert to be an accurate representation of the clinical information, it is an objective way of establishing the completeness, accuracy, and overall validity of the model.

A significant number of sentences in the evaluation could not be processed by the natural-language processor because the lexicon was not complete and also because of spelling errors. These problems are relatively straightforward to address. The natural-language processor was not designed to handle spelling errors, but the radiology reporting system plans on incorporating a commercial spell-checking facility, which should eliminate most of the spelling errors. Extension of the lexicon should also cause an increase in the success rate of the processor. Undefined words are routinely recorded by the natural-language processor so that they can be classified and added to the lexicon.

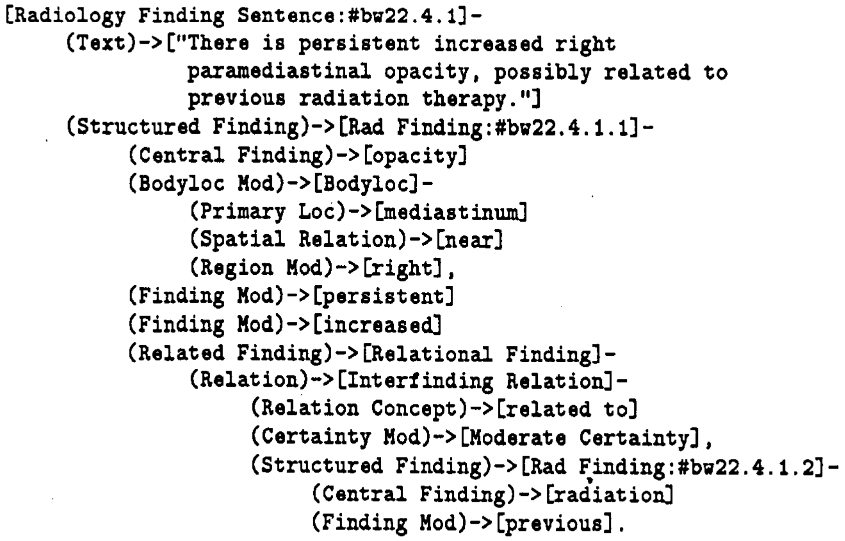
Our goal in encoding the information content is to obtain the degree of granularity needed for reliable retrieval by our automated decision support system. In addition, we want to facilitate the writing of the query that will be subsequently used to automatically retrieve the encoded data. The representation of the clinical information was therefore deliberately simplified as much as possible. Although the conceptual model can represent information nested to an arbitrary depth (i.e., qualifiers may have qualifiers), the natural-language processor generates limited values for these types of qualifiers. For the same reason, nested information is flattened as much as possible when the encoded data are uploaded to the coded patient database.

The following is a summary of information that could not be represented accurately by our schema because it has a simplified form:

1. Comparisons between finding qualifiers associated with different body locations cannot be precisely represented. In *right greater than left pleural effusions,* the model represents a right pleural effusion and a left pleural effusion, but cannot represent that the effusion on the left is greater than the one on the right. A similar problem exists for *markings are more scattered on the right than on the left* and for *the mediastinal wire staples are positioned more medially in relation to the mediastinum on the right.* Notice that the comparison relation is between the qualifiers of the findings and not the findings. We do not believe that this problem will have an impact on decision support because presently it is sufficient to test for the presence or absence of a finding (e.g., pleural effusion). Temporal comparisons, which occur more frequently, such as *increased interstitial markings* and *heart is more transversely positioned,* are represented appropriately.

2. Body location information is coarsely represented. It is not intended to be at the level of granularity whereby the body location of the finding is precisely pinpointed. Presently, for decision support, it is enough to know that a finding is or is not present in a certain gross body location; whether it is present in the upper part, lower part, or base of that location is not relevant. Similarly, for a sentence such as *opacity between 1st and 2nd rib,* the model represents that there is an opacity associated with the two body location qualifiers *near 1st rib* and *near 2nd rib.*

3. Temporal information is simplified considerably in order to facilitate retrieval for decision support applications. For example, if the incidence of a finding has increased from a previous examination, the information is captured simply by associating a qualifier **increase** with the finding. Therefore, writing a query associated with this type of temporal information is relatively easy.



```
[Radiology Finding Sentence:#bw22.4.1]-
    (Text)->["There is persistent increased right
              paramediastinal opacity, possibly related to
              previous radiation therapy."]
    (Structured Finding)->[Rad Finding:#bw22.4.1.1]-
        (Central Finding)->[opacity]
        (Bodyloc Mod)->[Bodyloc]-
            (Primary Loc)->[mediastinum]
            (Spatial Relation)->[near]
            (Region Mod)->[right],
        (Finding Mod)->[persistent]
        (Finding Mod)->[increased]
        (Related Finding)->[Relational Finding]-
            (Relation)->[Interfinding Relation]-
                (Relation Concept)->[related to]
                (Certainty Mod)->[Moderate Certainty],
                (Structured Finding)->[Rad Finding:#bw22.4.1.2]-
                    (Central Finding)->[radiation]
                    (Finding Mod)->[previous].
```

Figure 9 The structured output form of a sentence from a report after the schema has been applied to the sentence. The sentence has a primary finding **opacity**, which is qualified by a body location **mediastinum**. **Mediastinum** has a qualifier **right** and a spatial relation **near** denoting that the opacity is not contiguous to the mediastinum. The sentence also has a secondary finding **radiation** that is qualified by a temporal modifier **previous**. The secondary finding is connected to the primary finding via the relation **related to**, which is associated with a certainty modifier **Moderate Certainty**.

4. Values for degree, certainty, status, and temporal qualifiers are limited. For example, the degree phrase *severe* would be translated into a degree qualifier with the value **high degree**. Similarly, *possible* will be translated into a certainty qualifier with the value **moderate certainty**. Some information is lost in the process, but we do not believe this will have a detrimental impact on decision support since these qualifiers are generally vague and inaccurate at best[15] and are valid for decision support using only a coarse level of granularity.

5. When the encoded form is uploaded to the central patient database, nested qualifier information is flattened further. For example, after natural-language processing, the encoded representation of *very slight* will have a degree qualifier that has the value **low degree** (corresponding to *slight*) with a nested degree qualifier **high degree** (corresponding to *very*). When this form is uploaded to the database, the degree qualifier is not nested further but is represented simply as having the value **low degree**.

6. Relations among findings are also simplified when uploading the encoded form to the coded patient database in order to eliminate the nesting of findings. Thus, if the sentence contains an inference, as in *new opacities may represent consolidation*, the structured output will be translated into two separate findings, *new opacities* and *possible consolidation*.

The schema presented in this paper will continue to evolve as more and more reports are automatically analyzed by the natural-language processor and the results of the processing are evaluated further. Changes to the schema will be made when relevant clinical information that is needed by the decision support system occurs in the reports but cannot be mapped into the conceptual model because it cannot adequately represent the information. Another reason the schema may be modified is to facilitate the writing of automated queries for decision support applications.

Other types of information that may need further work are associated with the representation of various types of modifier information. Certain types of modifiers were deliberately represented in a simplified form in order to facilitate the writing of automated queries. For example, temporal information is represented as a modifier of a finding. Writing queries associated with this simple schema is still quite complicated because the modifiers significantly affect the interpretation of the findings being retrieved. For example, if a decision support system is retrieving reports that may imply **congestive heart failure**, it is appropriate to select reports containing a finding **congestive heart failure** (along with other finding concepts such as **pleural effusion** associated with congestive heart failure) only under the following circumstances:

1. The temporal modifier is not **resolved** or **decreased**.
2. The certainty modifier is not **no** or **undetermined**.
3. The degree modifier is not **low degree**.

The trade-off between expressiveness and practicality is a very fine one that will have to be continuously balanced.

Another issue concerning the schema is whether it can be extended to other domains. The formalism of the CG schema itself is general, as is the formalization of the linguistic properties of medical language. The details and granularity may vary among domains and applications, but the methodology itself is general and does not depend on properties specific to radiology reports. Therefore, it should be possible in principle to develop the same or similar schemas for other domains and applications. However, the balancing of trade-offs will constantly have to be addressed and evaluated anew because most likely there is no optimal granularity for all uses and for all domains.

The natural-language processor, including the grammar, lexicon, compositional mappings, and synonym knowledge base, is implemented for this domain, and has been used to automatically structure the information in the radiology reports in accordance with the schema. The controlled vocabulary and the structure of the information is presently being integrated with the production version of the MED for use in the production system.

Conclusions

The results presented here were derived from a combination of methods, some manual and some automated. The approach is one that combines a top-down view of analysis of the reports and a bottom-up view to identify the basic concepts and the various ways of expressing them in the reports. The resulting structures represent radiology findings in a way that is a) derivable through automated natural-language processing, b) consistent with the form required for the MED, c) consistent with the storage requirements of the clinical database, and d) usable by automated decision support.

The authors thank Dr. Stanley Huff, from LDS Hospital, Drs. Robert Greenes and Douglas Bell, from Brigham and Women's Hospital, and Dr. Bill Hersh, from Oregon Health Sciences University, for providing radiology reports of the chest from their institutions to enrich the collection of reports from CPMC; Drs. Philip Alderson and John Austin, from the Radiology Department, CPMC, for providing their expertise and help in the interpretation of the radiology reports and in the manual evaluation of the structured forms; and Dr. George Hripcsak, from the Center for Medical Informatics, Columbia University, for providing advice relating to the interaction between the schema and the writing of automated queries for the decision support system.

References ■

1. Levesque HJ, Brachman RJ. A fundamental tradeoff in knowledge representation and reasoning. In: Readings in Knowledge Representation. San Mateo, CA: Morgan Kaufmann, 1985:42–70.
2. Hripscak G, Cimino JJ, Johnson SB, Clayton PD. The Columbia-Presbyterian Medical Center decision-support system as a model for implementing the Arden Syntax. In: Clayton PD, ed. Fifteenth Annual Symposium on Computer Applications in Medical Care, Washington, DC, November 17–20, 1991. New York: McGraw-Hill, 1992:248–52.
3. Cimino JJ, Clayton PD, Hripcsak G, Johnson SB. Knowledge-based approaches to the maintenance of a large controlled medical terminology. J Am Med Informatics Assoc. 1994;1(1):35–50.
4. Friedman C, Alderson PO, Austin JHM, Cimino JJ, Johnson SB. A general natural language text processor for clinical radiology. J Am Med Informatics Assoc. 1994;1(2):161–74.
5. Friedman C, Cimino JJ, Johnson SB. A conceptual model for clinical radiology reports. In: Safran C, ed. Seventeenth Symposium for Computer Applications in Medical Care. New York: McGraw-Hill, 1993:829–33.
6. Masarie FE Jr, Miller RA, Bouhaddou O, Giuse NB, Warner HR. An interlingua for electronic interchange of medical information: using frames to map between clinical vocabularies. Comput Biomed Res. 1991;24:379–400.
7. Rector AL, Nowlan WA, Kay S. Conceptual knowledge: the core of medical information systems. In: Lun KC, Degoulet P, Plemme TE, Rienhoff O, eds. Proceedings of MEDINFO 92. Amsterdam: North-Holland, 1992:1420–6.
8. Evans DA. Final report on the MedSORT-II project: developing and managing medical thesauri. Technical Report no. CMU-LCL-87-3. Pittsburgh: Laboratory for Computational Linguistics, Carnegie Mellon University, 1987.
9. Rossi-Mori. CEN/TC251/PT003 model for representation of terminologies and coding systems in medicine. In: Proceedings of the Seminar: Opportunities for European and US Cooperation in Standardization in Health Care Informatics, Geneva, Switzerland, September 1992.
10. Côté RA, Rothwell DJ, Palotay JL, Beckett RS, Brochu L, eds. The systematized nomenclature of medicine: SNOMED international. Northfield, IL: College of American Pathologists, 1993.
11. Sager N, Friedman C, Lyman MS, et al. Medical Language Processing: Computer Management of Narrative Data. Reading, MA: Addison-Wesley, 1987.
12. Sowa JF. Conceptual Structures. Reading, MA: Addison-Wesley, 1984.
13. Friedman C, Hripcsak G, Johnson SB, Cimino JJ, Clayton PD. A generalized relational schema for an integrated clinical patient database. In: Miller RA, ed. Proceedings of the Fourteenth Annual Symposium on Computer Applications in Medical Care, Washington, DC, November 4–7, 1990. Los Alamitos, CA: IEEE Computer Society Press, 1990:335–9.
14. Johnson SB, Friedman C, Cimino JJ, Clark AS, Hripcsak G, Clayton PD. A conceptual schema for a central patient database. In: Clayton PD, ed. Proceedings of the Fifteenth Annual Symposium on Computer Applications in Medical Care, Washington, DC, November 17–20, 1991. New York: McGraw Hill, 1992:381–7.
15. Nakao MA. Numbers are better than words: verbal specifications of frequency have no place in medicine. Am J Med. 1983;74:1061–4.

APPENDIX A
Structured X-ray Reports

```
Sentence bw22.1.1 was not processed because it does not contain any findings.
It is "PA view is compared to the previous examination dated 10-22-91."

[Radiology Finding Sentence:#bw22.2.1]-
     (Text)->["Surgical clips are again seen along the
                right mediastinum and right hilar region."]
     (Structured Finding)->[Rad Finding:#bw22.2.1.1]-
               (Central Finding)->[surgical clip]
               (Bodyloc Mod)->[Bodyloc]-
                    (Primary Loc)->[mediastinum]
                    (Region Mod)->[right],
               (Bodyloc Mod)->[Bodyloc]-
                    (Primary Loc)->[lung]
                    (Region Mod)->[hilum]
                    (Region Mod)->[right],
               (Finding Mod)->['>1']
               (Finding Mod)->[high certainty]
               (Finding Mod)->[again].

[Radiology Finding Sentence:#bw22.3.1]-
     (Text)->["There are new surgical clips in the
        distribution of the circumflex artery as well as 4 intact
        sternotomy wires."]
     (Structured Finding:#bw22.3.1.1)->[Rad Finding:#bw22.3.1.1]-
              (Central Finding)->[surgical clip]
              (Bodyloc Mod)->[distribution of circumflex artery]
              (Finding Mod)->['>1']
              (Finding Mod)->[new],
     (Structured Finding:#bw22.3.1.2)->[Rad Finding:#bw22.3.1.2]-
               (Central Finding)->[sternotomy wire]
               (Finding Mod)->[intact]
               (Finding Mod)->[4]
               (Finding Mod)->['>1'].

[Radiology Finding Sentence:#bw22.4.1]-
     (Text)->["There is persistent increased right
                 paramediastinal opacity, possibly related to
                 previous radiation therapy."]
     (Structured Finding)->[Rad Finding:#bw22.4.1.1]-
              (Central Finding)->[opacity]
              (Bodyloc Mod)->[Bodyloc]-
                   (Primary Loc)->[mediastinum]
                   (Spatial Relation)->[near]
                   (Region Mod)->[right],
              (Finding Mod)->[persistent]
              (Finding Mod)->[increased]
              (Related Finding)->[Relational Finding]-
                   (Relation)->[Interfinding Relation]-
                       (Relation Concept)->[related to]
                       (Certainty Mod)->[Moderate Certainty],
                   (Structured Finding)->[Rad Finding:#bw22.4.1.2]-
                       (Central Finding)->[radiation]
                       (Finding Mod)->[previous].

[Radiology Finding Sentence:#bw22.5.1]-
     (Text)->["New platelike opacities are seen in the
```

```
        left and mid lower lung zones, compatible with atelectasis."]
    (Structured Finding)->[Rad Finding:#bw22.5.1.1]-
            (Central Finding)->[opacity]
            (Bodyloc Mod)->[BodyLoc]-
                (Primary Loc)->[lung]
                (Region Mod)->[mid]
                (Region Mod)->[lower],
            (Region Mod)->[left]
            (Finding Mod)->[new]
            (Finding Mod)->[platelike]
            (Finding Mod)->[high certainty]
            (Related Finding)->[Relational Finding]-
                (Relation)->[compatible with]
                (Structured Finding)->[Rad Finding:#bw22.5.1.2]-
                   (Central Finding)->[atelectasis].

[Radiology Finding Sentence:#bw22.6.1]-
    (Text)->["There has been some interval improvement
           in the left pleural effusion."]
    (Structured Finding)->[Rad Finding:#bw22.6.1.1]-
            (Central Finding)->[pleural effusion]
            (Region Mod)->[left]
            (Finding Mod)->[Temporal]-
                (Temporal Concept)->[improvement]
                (Degree Mod)->[low degree].

[Radiology Finding Sentence:#bw22.7.1]-
    (Text)->["Slight interval decrease in left pleural
             effusion."]
    (Structured Finding)->[Rad Finding:#bw22.7.1.1]-
            (Central Finding)->[pleural effusion]
            (Region Mod)->[left]
            (Finding Mod)->[Change Mod]-
                (Change Concept)->[decrease]
                (Degree Mod)->[low degree].

[Radiology Finding Sentence:#bw22.8.1]-
    (Text)->["Left lower lobe atelectasis."]
    (Structured Finding)->[Rad Finding:#bw22.8.1.1]-
            (Central Finding)->[atelectasis]
            (Bodyloc Mod)->[Bodyloc]-
                (Primary Loc)->[Left Lower Lobe of Lung].

[Radiology Finding Sentence:#bw22.9.1]-
    (Text)->["Post - operative changes consistent with
          coronary artery bypass graft as well as previous
          lobectomy on the right."]
    (Structured Finding)->[Rad Finding:bw22.9.1.1]-
            (Finding Mod)->[postoperative change]
            (Related Finding)->[Relational Finding]-
               (Relation)->[consistent with]
               (Structured Finding)->[Rad Finding:#bw22.9.1.2]-
                  (Evidential Procedure)->[coronary artery
                     bypass graft],
               (Structured Finding)->[Rad Finding:#bw22.9.1.3]-
                  (Region Mod)->[right]
                  (Evidential Procedure)->[lobectomy]
                  (Finding Mod)->[previous].
```

```
[Radiology Finding Sentence:#lds71.1.1]-
    (Text)->["In comparison to a study of 6 / 2 / 92, there
              has been a slight increase in the degree of cardiomegaly."]
    (Structured Finding)->[Rad Finding:#lds71.1.1.1]-
            (Central Finding)->[cardiomegaly]
            (Finding Mod)->[Temporal]-
                (Temporal Concept)->[increase]
                (Degree Mod)->[low].

[Comp Report:[date:6:2:92]]

[Radiology Finding Sentence:#lds71.2.1]-
    (Text)->["Pulmonary vessels remain distended."]
    (Structured Finding)->[Rad Finding:#lds71.2.1.1]-
             (Central Finding)->[Distended]
             (Bodyloc Mod)->[Pulmonary Vessel]
             (Finding Mod)->[remain]
             (Finding Mod)->['>1']]

[Radiology Finding Sentence:#lds71.4.1]-
    (Text)->["Pleural fluid is tracking into the
              fissures."]
    (Structured Finding)->[Rad Finding:#lds71.4.1.1]-
             (Central Finding)->[Pleural Fluid]
             (Bodyloc Mod)->[Bodyloc]-
                 (Primary Loc)->[fissure]
                 (Spatial Mod)->[track into].

[Radiology Finding Sentence:#lds71.5.1]-
    (Text)->["There is minimal blunting of the
              costophrenic angles."]
    (Structured Finding)->[Rad Finding:#lds71.5.1.1]-
              (Central Finding)->[blunting]
              (Bodyloc Mod)->[costophrenic angle]
              (Finding Mod)->[low degree].

[Radiology Finding Sentence:#lds71.6.1]-
    (Text)->["Slight increase in the degree of
              congestive heart failure."]
    (Structured Finding)->[Rad Finding:#lds71.6.1.1]-
              (Central Finding)->[congestive heart failure]
              (Finding Mod)->[Temporal]-
                  (Temporal Concept)->[increase]
                  (Degree Mod)->[low degree].

[Radiology Finding Sentence:#lds71.7.1]-
    (Text)->["New opacities in the right base
              consistent with consolidation."]
    (Structured Finding)->[Rad Finding:#lds71.7.1.1]-
              (Central Finding)->[opacity]
              (Region Mod)->[right]
              (Region Mod)->[base]
              (Finding Mod)->[new]
              (Finding Mod)->['>1']
              (Related Finding)->[Relational Finding]-
                  (Relation)->[consistent with]
                  (Structured Finding)->[Rad Finding:#lds71.7.1.2]-
                       (Central Finding)->[consolidation].
```

Correction

Author Carol Friedman's affiliation was incomplete in the footnote to her article in the March/April issue (Friedman C, Cimino JJ, Johnson SB, A schema for representing medical language applied to clinical radiology, JAMIA 1994;1:233–248). Her major affiliation, with Queens College of the City of New York, was inadvertently omitted.