Research Paper ■

A General Natural-language Text Processor for Clinical Radiology

CAROL FRIEDMAN, PHD, PHILIP O. ALDERSON, MD, JOHN H. M. AUSTIN, MD, JAMES J. CIMINO, MD, STEPHEN B. JOHNSON, PHD

Abstract **Objective**: Development of a general natural-language processor that identifies clinical information in narrative reports and maps that information into a structured representation containing clinical terms.

Design: The natural-language processor provides three phases of processing, all of which are driven by different knowledge sources. The first phase performs the parsing. It identifies the structure of the text through use of a grammar that defines semantic patterns and a target form. The second phase, regularization, standardizes the terms in the initial target structure via a compositional mapping of multi-word phrases. The third phase, encoding, maps the terms to a controlled vocabulary. Radiology is the test domain for the processor and the target structure is a formal model for representing clinical information in that domain.

Measurements: The impression sections of 230 radiology reports were encoded by the processor. Results of an automated query of the resultant database for the occurrences of four diseases were compared with the analysis of a panel of three physicians to determine recall and precision.

Results: Without training specific to the four diseases, recall and precision of the system (combined effect of the processor and query generator) were 70% and 87%. Training of the query component increased recall to 85% without changing precision.

■ **J Am Med Informatics Assoc.** 1994;1:161–174.

Natural language is the most widespread, comprehensive, and convenient medium in which health care personnel can express clinical information. It is not, however, suitable for important computerized applications such as automated quality assurance, clinical decision support, and research, which require error-free access to clinical information. Within these applications, access is typically achieved by limiting data entry to a controlled vocabulary consisting of a set of unique, well-defined, unambiguous medical concepts that generally do not correspond directly to the clinical information in patient documents. Although presently there is a steadily increasing supply of clinical data available in electronic form, a large portion of the information remains inaccessible because it is in the form of narrative text. Unfortunately, there is a large gap between terminology expressed as controlled vocabulary and clinical information as expressed naturally in most texts containing patient data.

Several major developments have contributed to an increased need for computerized methods that process clinical information expressed in the form of natural language:

- The capture of data in electronic form is becoming commonplace. Much clinical information that is maintained online is typically in the form of free text (i.e., procedure reports, history, progress notes, discharge summaries, operative notes, physical examination, and admission summaries).

Affiliations of the authors: Columbia University (CF, JJC, SBJ) and Columbia Presbyterian Medical Center (POA, JHMA), New York, NY.

Supported in part by Grant Number R29 LM05397 from the National Library of Medicine and Grant Number 6-61483 from the Research Foundation of CUNY.

Correspondence and reprints: Carol Friedman, PhD, Queens College of the City University of New York, Computer Science Department, Flushing, NY 11367–1597.

Received for publication: 7/28/93; accepted for publication: 11/11/93.

- Automated clinical systems that depend on the availability of coded clinical data are being used increasingly in attempts to improve the quality of patient care.
- Advances in computer technology provide the capability to store huge amounts of data at reasonable cost and to process documents within a reasonable time frame.

In this paper we present a medical text processor that translates clinical information in patient documents into controlled vocabulary terms. Initially we have limited the domain to clinical radiology and the controlled vocabulary to concepts that are useful for decision support and research. The methodology, however, allows the domain and capabilities of the system to be extended in a modular and systematic manner, without being limited by and without having to change significantly the underlying approach.

Background

There are various ways in which coded clinical data can be obtained. Some types of clinical data are reasonably simple to obtain directly in coded form, such as clinical laboratory data, pharmacy orders, vital signs, and identification of clinical procedures. However, other types of data, such as findings from examinations, history, progress notes, and discharge summaries, are more elusive. Some systems rely on direct physician entry of coded data via graphic or form-based user interfaces,[1–3] and other systems use speech-recognition interfaces.[4–8] These systems all impose limitations on what the user can enter, although some have a comment field (which is not encoded) that captures information outside the system. Although these systems have many advantages, they are not as widely used as natural language.

Another limitation to direct coded data entry is that while simple information is reasonable to codify, complex information is time-consuming and difficult to code accurately. Consider the effort involved in representing the following simple phrases with modifiers using a predefined format: *decreasing but persistent pneumonia, no definite evidence of pneumonia, marked worsening of pneumonia, rule out acute pneumonia,* and *cannot definitely exclude pneumonia.*

Other methods of obtaining coded data rely on processing the text. One method is a pattern-matching technique that is a variation of keyword search. It is being applied in pathology reports,[9] discharge summaries,[10] and other domains.[11,12] This method is generally useful for texts that are naturally highly structured. For less structured text (e.g., descriptive sections of pathology and radiology reports, history, admission notes, progress reports) the same information is usually expressed in so many different ways, encompassing a large variety of stylistic linguistic variations, that it would be virtually impossible to enumerate them. However, when an expression of the text completely matches a pattern in the system, this method is both efficient and reliable. Yet, if some information in the text is skipped and only parts of the text match a pattern in the system, a serious misinterpretation may occur because neither semantic nor syntactic relations between words are recognized; e.g., compare the meanings of the different phrases containing *pain*: *severe pain, no relief of pain, pain continued, pain decreased slightly,* and *free of pain.* If certain parts of the phrases in the above examples were ignored, their meanings would be seriously misinterpreted.

Another approach combines concept-based matching algorithms[13] (developed for information retrieval applications) with restricted natural-language processing techniques to build applications that have limited, yet important goals. Applications using this technique have been built in the domains of physical examination findings and chest radiology reports.[14,15] In the domain of radiology reports, the efforts were focused on identifying patients who needed follow-up because their reports contained findings associated with potentially malignant lesions. In the domain of physical examinations, target findings specified by the user were identified. Natural-language processing in this system consists of using heuristic rules to break up the sentences into a series of phrases, then using finite-state automata to process the phrases. This system was shown to be very effective when the goals are restricted and the text highly structured.

Other techniques are based on semantic knowledge of the domain. Typically this method involves the development of a semantic representational model[16–20] that is in the form of frames[21] or conceptual graphs.[22] These forms describe meaningful predefined relations between semantic categories in the domain that occur in the texts. Text processing consists of mapping words and phrases in the text into the representational model based on the semantic classification of the phrases. This method is more general than the pattern-matching method because individual patterns do not have to be enumerated. Instead, a dictionary or lexicon is used to semantically classify words and phrases in the domain, and these classes are used to drive the mapping. This method also is effective for text that is naturally very structured. However, the absence of syntactic information

causes performance problems similar to those of the pattern-matching systems, particularly when the text is not highly structured. Complex language structures, such as embedded clauses, are difficult to handle adequately with this approach. In addition, if portions of the text are skipped, serious inaccuracies could occur. This method has been used to process echocardiographic findings,[19,23] discharge summaries,[18] and radiology findings.[20]

Another approach, which is more complex, requires both syntactic and semantic knowledge, and therefore it can handle a larger variety of linguistic structures. This technique has been developed by the Linguistic String Project (LSP)[24–26] and was used to process reports in several clinical domains, such as discharge summaries, progress notes, and radiology reports. In the LSP system, the natural-language expressions are mapped to structured predefined formats that represent the underlying concepts and relations in the domain. The values stored in the formats represent the regularized form of the linguistic expressions, but there is no mechanism that maps these forms to unique codes or to controlled vocabulary concepts. Systems containing syntactic knowledge are very time-consuming to build and maintain because syntax is so complex. In addition, they are fragile because an undefined word, an unusual syntactic structure, or a new semantic pattern can cause a failure. However, they are generally accurate when sentences are successfully processed.

Methods

Overview

The text-processing system presented in this paper is semantically based, but some syntax is included to handle coordinate conjunctions, such as *and* and *or*, and simple relative clauses. The system contains a parser that determines the structure of the text. The parser is driven by a semantic grammar that is highly effective for handling structured text and common patterns. Therefore, the system is very suitable for the domain of radiology.

The grammar consists of rules specifying well-defined semantic patterns, their interpretations, and the underlying target structures into which they should be mapped. In our system, the target structures correspond to the formal model of the domain. The semantic grammar incorporates pattern-matching and semantic techniques into one formalism (the semantic grammar), but is more general. Our system differs from the other semantic-based systems described above because the text is analyzed and structured by following the grammar rules exactly. High accuracy is achieved because the parser is constrained so that it is successful only if the sentence corresponds to one of the well-formed semantic patterns specified in the grammar. Generally, if a well-formed semantic pattern is found, it reflects directly the underlying semantic relations among concepts in the domain, and a correct interpretation and translation to the target structure are highly likely. This is a strong claim. It means that the ambiguities generally present in natural language as a whole are reduced markedly within the language of the domain because of the underlying semantics.

An example of a semantic pattern is DEGREE + CHANGE + FINDING, which consists of degree information followed by change and finding information, as in *mild increase in congestion*. In this example, *mild* is associated with degree information, *increase in* with change information, and *congestion* with finding information. This pattern is interpreted so that DEGREE qualifies CHANGE, which together qualifies FINDING. In this case, *mild* qualifies *increase*, and *mild increase* qualifies *congestion*. If the sentence were *mildly increased congestion* the semantic interpretation would be exactly the same, although the syntactic structures of the phrases are somewhat different. This means that the phrases *mildly increase in congestion* and *mild increase congestion* would also be acceptable, although syntactically they are not correct. However, we have found that clinical documents contain many syntactic structures that are considered incorrect according to general English grammar, and that the semantic pattern and not the syntactic pattern strongly determines the underlying interpretation.

Sometimes there are two or more possible interpretations for a particular semantic pattern. For example, in the phrase *mild increased congestion*, the qualifiers *mild* and *increased* may both modify *congestion*, but this interpretation of the co-occurrence pattern is less frequent, and therefore the interpetation where *mild* modifies *increased* is chosen.

Figure 1 shows a schematic overview of the processor, which consists of three phases of processing. The first stage of the processing contains the parser, which determines the structure of the text and generates the preliminary structured output form for the clinical information in the text. The parser uses a grammar and lexicon to determine the structure of the text and to translate it to the target form. The structuring of text is a critical and difficult step in the processing and results in a great reduction of the stylistic variations found in natural-language expressions. For example, the target forms for many phrases

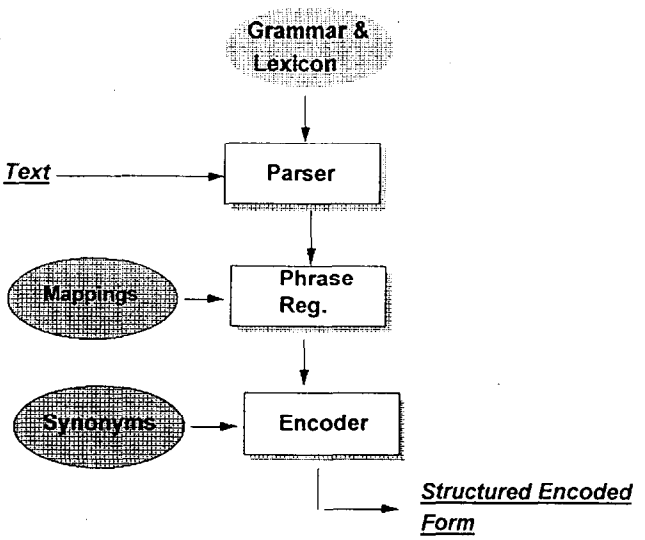

Figure 1 The schemata of the text processor.

that are lexical variants, such as *enlarged heart, cardiac enlargement,* and *enlargement of the heart* will be the same as a result of the structuring process. However, not all variations are reduced to one form by this stage of processing, and the structured forms do not yet correspond to unique controlled vocabulary concepts.

The second stage of processing is the phrase-regularization phase, which is required to further reduce stylistic variations that occur in natural language. This is considered a compositional component, because in this phase, the structured outputs of noncontiguous expressions that occur in the reports are combined and standardized so that they correspond to the appropriate regularized forms. This is accomplished using a mapping knowledge base, which consists of the structured output forms of multi-word phrases that can be decomposed. For example, the phrase *heart appears to be slightly enlarged* is a variant form of *enlarged heart* with modifiers. For this sentence, the processor will initially produce a structured output form where the value of the central finding corresponds to *enlarged*. The central finding will also have a body location modifier whose value corresponds to *heart*, and degree and certainty modifiers corresponding to *appears* and *slightly*. This component will find that the output form for the above sentence subsumes the output form of the phrase *enlarged heart* because it also consists of a central finding corresponding to *enlarged* and a body-location modifier corresponding to *heart*. Therefore the central finding from *heart appears to be slightly enlarged* will be changed from *enlarged* to *enlarged heart*, but the body location, degree, and certainty modifiers will remain the same.

Once the contiguous and non-contiguous lexical variants have been mapped to standard forms, the last task is the encoding phase, which maps the standard forms into unique concepts associated with the controlled vocabulary. This is a straightforward task involving a one-to-one mapping, because all the linguistically variant forms have already been reduced to standardized target forms. This mapping is accomplished by means of a synonym knowledge base that consists of standard forms and their corresponding concepts in the controlled vocabulary. Thus, this knowledge base forms a critical bridge between the language of the text and the unique concepts in the controlled vocabulary.

The Medical Entities Dictionary (MED) developed at Columbia Presbyterian Medical Center (CPMC)[27] is a knowledge base of medical concepts that consist of taxonomic relations in addition to other relevant semantic relations. Using the synonym knowledge base, the regularized forms are translated into unique concepts so that when the final structured forms of the processed reports are uploaded to the centralized patient database,[28,29] they correspond to unique concepts in the MED. All applications at CPMC, such as decision support, that require reliable access to coded clinical data can then reliably access the data by queries that utilize the structured form and the controlled vocabulary of the MED.

The Formal Representation of Clinical Radiology

In order to map the clinical information in the patient documents into a structured form, a formal model was designed to represent the clinically salient information. The fundamental design of this model is based on the information formats developed by the Linguistic String Project.[24] Others[30] have also developed models representing the informational content of clinical information in the domain of chest x-rays, in the domain of cardiopulmonary diseases,[31] and in the domain of general medical terminology.[32–34] This section contains a brief description of two of the most relevant components of the representational model, **Rad Finding Structure** and **Modifier**, which represent the structures of the findings and the modifiers, respectively. The representation of body-location information is not shown here. A more detailed description of the complete model, which also represents contextual information and body locations as well as findings, is given elsewhere.[35]

The structure of the simplified model for report findings and modifiers is shown in Figure 2 using the linear notation for Conceptual Graphs (CGs).[22] In

```
[Rad Finding Structure]-
     (Central Finding)->[Rad Finding:{*}]
     (Bodyloc Mod)->[Bodyloc:{*}]
     (Finding Mod)->[Modifier:{*}]

[Modifier]-
     (Certainty Mod)->[Certainty:{*}]
     (Degree Mod)->[Degree:{*}]
     (Change Mod)->[Change:{*}]
     (Status Mod)->[Status:{*}]
     (Quantity Mod)->[Quantity:{*}]
     (Descriptor Mod)->[Descriptor:{*}]
```

Figure 2 Representation of radiology findings and finding modifiers.

CGs a concept is similar to a frame, and a relation is similar to a slot. A concept* is enclosed in square brackets, 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 a *domain* concept that appears in square brackets after the arrow. Thus, the general format of a concept with N relations is:

[Concept]-
(Relation1)→[Domain1]
(Relation2)→[Domain2]
:
(RelationN)→[DomainN]

Findings consist typically of a central finding and modifiers that contain body-location information represented by the relation **Bodyloc Mod**, and other modifiers such as certainty, severity, and temporal qualifiers, represented by **Finding Mod**. The relation **Central Finding** associates the structured finding with a central finding. For example, in *severe scarring*, *scarring* would be the value of **Central Finding**. Even though Figure 2 specifies every relation associated with **Rad Finding Struct** as optional, the processor will never generate empty structures.

The representation of the modifier structure **Finding Mod** shown in Figure 2 consists of different informational types of modifiers. The relation **Certainty Mod** is associated with certainty information related to the finding. Because there are many words and phrases associated with this type of information, and because their underlying meanings are vague, the concepts in the controlled vocabulary associated with this type of information are limited to five concepts: **no**, **low certainty**, **moderate certainty**, **high certainty**, and **cannot evaluate**, and therefore the words and phrases in the reports relating to certainty information are mapped into one of the five appropriate concepts. This delimitation greatly facilitates the subsequent retrieval of the structured findings without a significant loss of precision. Because this type of information is mostly used in the reports to hedge information concerning the certainty of the findings and is basically very imprecise from the start, we have not found it is useful or accurate to represent this type of information more precisely. Concepts corresponding to other types of vague or qualitative information, represented by degree, change, and status modifiers, are also limited for the same reason. In other applications, it may be important to handle qualitative information more precisely, in which case a different representation may be desirable.

Radiology findings interact with each other when one finding is related to a second finding. The interactions are expressed in the text in the form of connective semantic relations such as *may represent*, *suggests*, *consistent with*, and *indicative of*. These relations generally denote interpretations and therefore also lack precision. For example, in *mildly increased interstitial markings may be indicative of pneumocystis or viral pneumonia*, the observation is *mildly increased interstitial markings* and its interpretation is that *pneumocystis* is possible or *viral pneumonia* is possible. To simplify retrieval of the structured information, connective information is not represented directly, because each finding is represented independently. However, the connective relation is included as a certainty modifier and is associated with the structured finding(s) that follow the connective relation. Thus, in *ill-defined left perihilar density may represent an infiltrate*, the structured form containing the finding *infiltrate* will have a certainty modifier corresponding to the concept **moderate certainty** which is denoted by *may represent*. In *mildly increased interstitial markings may represent pneumocystis or viral pneumonia*, the findings *pneumocystis* and *viral pneumonia* will both have certainty modifiers corresponding to the concept denoted by *may represent*. It could be argued that by simplifying the relations in this way, we are losing important connective information. However, we have found that this type of information is vague, and that including it greatly complicates retrieval. If we subsequently find that it is important, we will modify the representation and mappings accordingly. Parallel findings, such as *hyperinflated lungs with pleural effusion*, and *left lower lobe infiltrate and pleural effusion*, are

Notice that the main concept is followed by a dash (-) and is terminated by a period (.). The number of values that a relation is permitted to have (its cardinality) is indicated by including a constraint C following the domain name. If C is :{} the relation may have 0 or more values; if it is :@>1, the relation may have 1 or more values; if it is :@<2, the relation may have 0 or 1 values, and if it is :@1, the relation must have exactly 1 value.

```
<Sentence>     --> <Patterns> ("."| ";") .
<Patterns>     --> <FindingRel> {<MoreFinding>} .
<FindingRel>   --> <FindingPhr> |
                   <BpMods> <Verbrel> <FindingPhr> .
<FindingPhr>   --> {<Lmods>} <Findingterm> {<Rmods>}.
<MoreFinding>  --> <Relation> <FindingRel> .
<Findingterm>  --> disease | cfinding | pfinding | descriptor.
<Lmods>        --> [<CertMods>] [<DegreeMods>] [<ChangeMods>] [<BpMods>].
<Rmods>        --> <SpatialRel> <BpMods> | <Lmods> .
<CertMods>     --> [negation] certainty | negation.
<Relation>     --> conjunction | <Verbrel> .
<Verbrel>      --> [auxverb] [be] [negation] certainty.
<DegreeMods>   --> degree .
<ChangeMods>   --> [negation] change .
<BpMods>       --> {<RegionMods>} bodyloc {<MoreBpMods>}.
<MoreBpMods>   --> <SpatialRel> <BpMods> | conjunction <BpMods> .
<SpatialRel>   --> in | on | at | along | near | under .
<RegionMods>   --> region {<MoreRegion>}.
<MoreRegion>   --> conjunction <RegionMods>.
```

Figure 3 Simplified semantic grammar for radiology findings.

represented as independent findings and their certainty modifiers are not affected.

The Parser

The parsing component of the processor uses an extended context-free grammar[36,37] to parse text sentences and to translate them to structured forms. This is the first stage in the processing of the text, as shown in Figure 1. The extended grammar includes translation rules to translate the grammatical structures into target forms, which correspond to the model shown in Figure 2, and constraints to specify well-formedness restrictions for the grammatical structures. The grammar being described in this section is a semantic grammar and it delineates semantic relations and structures. It is presently written in a form interpretable directly by Prolog. A small simplified semantic grammar for this domain is shown in Figure 3 in extended Backus Naur form[36] with the translation rules omitted for ease of illustration. The actual semantic grammar used to process the text is much larger and contains approximately 350 grammar rules. A separate discussion of the translation rules is provided below.

In Figure 3, the symbols enclosed in angle brackets specify the names of non-atomic semantic structures that are also defined in the grammar, the symbols enclosed in quotes are literals, the symbols enclosed in square brackets are optional, the symbols enclosed in curly brackets may occur zero or more times, and plain symbols, such as **disease**, **certainty**, and **cfinding** correspond to atomic components that are the semantic categories directly associated with single words or multiword phrases of the sentences being processed. According to the grammar in Figure 3, a sentence consists of semantic patterns **Patterns** followed by an end mark, which is either a period (.) or a semicolon (;). **Patterns** consists of one finding relation called **FindingRel** optionally followed by more finding relations **MoreFinding**, as in *interstitial markings may be suggestive of edema* and *pleural effusion as well as congestion*. **FindingRel** consists of either a finding phrase **FindingPhr**, which is a single finding **Findingterm** with optional left and right modifiers (**Lmods** and **Rmods**), as in *mild chronic pleural effusions*, or with a body location structure called **BpMods** related to **FindingPhr** by a verb relation **Verbrel**, as in *heart appears slightly enlarged* or *heart is not enlarged*. **Findingterm** could be a word that has the semantic classification **disease** or it could be a word that is associated with the various other semantic finding classifications **cfinding**, **pfinding**, or **descriptor**. A description of the semantic classifications is given in Table 1 along with examples. **Lmods** and **Rmods** are modifiers of the finding and consist of negation, certainty, degree, change, and body location types of information.

Negation, which presents a particularly troublesome aspect of natural-language processing, is handled by the semantic grammar. Negation is specified in the grammar as an atomic category **negation**. This permits the grammar to cover semantic patterns consisting of negation followed by certainty, change or finding information, as in *no evidence of edema, no change in edema*, or *no edema*, and also to specify patterns where the verb is negated, as in *heart is not enlarged*. The target structure for negation is a finding qualifier **Certainty Mod** whose value is **no**. However, a special post-processing clean-up procedure, which is discussed below, is used to resolve inconsistencies in the structured output that are produced when a sen-

tence contains negation along with other certainty information.

The semantic lexicon is a separate component of the semantic grammar. It semantically classifies both single words and multi-word phrases and specifies their canonical forms. Our lexicon currently contains about 1,720 single-word entries and 1,400 multi-word phrases. The semantic classes in the lexicon correspond to the atomic components in the semantic grammar. Some examples of lexical entries are shown in Figure 4. Single words are classified by using the symbol **word**, and multi-word phrases are classified by using the symbol **phrase** or **aphrase**.

In Figure 4, the word *cardiac* is assigned the semantic class **bodyloc** (which is used for words denoting specific body locations) and the target form **heart** (which is the English canonical form that corresponds to the Greek term). The word *cardiomegaly* is assigned the semantic class **cfinding** (which is used for words that implicitly denote body location information in addition to the finding) and the corresponding target form **cardiomegaly**. Similarly, the target form for the word *enlarged* is **enlarged**, and its semantic class is **pfinding** because it corresponds to a finding where the body location information is unknown. The semantic classification and target form for *enlargement* are the same. The semantic class for *improved* is **change** and the target form is **improve**. The classification and target forms for *improve*, *improvement*, and *improving* are also the same. In most cases, the specification of the target form is the form of the word that occurs most frequently in the domain. For example, *enlarged* occurs more often than *enlargement* and *enlarge*, and therefore it is considered the target form.

Multi-word phrases consisting of more than one word are treated as single entities by the parser. A phrase is entered in the lexicon because it is either more efficient or more precise to handle as a single unit than as a combination of independent components. The lexical entry for a phrase is similar to the entry for a single word except the first argument is the first word in the phrase, which is used for indexing. The second argument is the semantic category, the third argument is a list comprising the individual words of the phrase, and the fourth argument specifies the target form, which is a character string. For example, the lexical entry corresponding to the phrase *enlarged heart* has the semantic class **cfinding** and the target form **enlarged heart**. The entry for *could not be evaluated*, is assigned the class **certainty** and the target form **cannot evaluate**.

The semantic lexicon also consists of a component that specifies those common words and phrases in

Table 1 ■

A Description of the Semantic Classes

Bodyloc	Terms denoting a well-defined area of the body or a body part. Examples: *hilum, left lower lobe, carotid artery*
Certainty	Terms affecting the certainty of a finding. This class modifies status and change terms in addition to findings. Examples: *possible, appears, no evidence of*
Cfinding	Terms denoting a complete radiology finding because these terms implicitly or explicitly contain a finding and a body location. Examples: *cardiomegaly, widening of mediastinum, pleural effusion*
Change	Terms denoting a change in findings where the change is an improvement or worsening of a finding but not the start or end. Examples: *worsening, improving, increase*
Connector	Terms that connect one finding to another. Examples: *may represent, indicative of, suggests*
Degree	Terms denoting the severity of a finding. These terms can also modify change, certainty, and other degree words. Examples: *mild, severe, moderate*
Descriptor	Terms qualifying a property of a body location or a finding. Examples: *linear, large, enlarged*
Device	Terms denoting surgical devices that are evident on the radiology report. Examples: *sternotomy wire, swan ganz catheter, surgical wires*
Disease	Terms denoting a disease. These terms are based on the disease axis in SNOMED3. Examples: *asthma, cardiomyopathy, sickle-cell disease*
Position	Terms denoting orientation. Examples: *transverse, anteroposterior, lateral*
Pfinding	Terms denoting a partial finding. These terms must occur along with a body location to be a complete finding. Examples: *opacity, lesion, markings*
Procedure	Terms denoting a therapeutic or diagnostic procedure. Examples: *bronchoscopy, mastectomy, radiation therapy*
Quantity	Terms representing non-numeric quantitative information. Examples: *many, few, multiple*
Recommend	Terms denoting recommendations. Examples: *clinical correlation, follow up, repeat xray*
Region	Terms denoting relative locations within a body location. Examples: *upper, lower, mid*
Status	Terms denoting temporal information other than an improvement or worsening of a finding. Examples: *chronic, active, resolved*
Technique	Terms denoting information related to the manner in which the radiographic examination was obtained. Examples: *expiratory film, poor inspiration*

```
word(cardiac,bodyloc,heart).
word(cardiomegaly,cfinding,cardiomegaly).
aphrase(could,certainty,[could,not,be,evaluated],'cannot evaluate').
word(enlarged,pfinding,enlarged).
phrase(enlarged,cfinding,[enlarged,heart],'enlarged heart').
word(enlargement,pfinding,enlarged).
word(heart,bodyloc,heart).
phrase(hilar,cfinding,[hilar,adenopathy],'hilar adenopathy').
word(improved,change,improve).
word(improvement,change,improve).
word(lung,bodyloc,lung).
word(mild,degree,mild).
word(pulmonary,bodyloc,lung).
aphrase(spinal,bodyloc,[spinal,canal],'spinal canal').
aphrase(swan,device,[swan,ganz,catheter],'catheter,Swan-Ganz').
aphrase(thoracic,bodyloc,[thoracic,aorta],'thoracic aortic').
phrase(tortuous,cfinding,[tortuous,aorta],'tortuous aorta').
phrase(uncoiled,cfinding,[uncoiled,aorta],'uncoiled aorta').
```

Figure 4 Single-word and phrasal lexical entries.

the domain that may be ignored in the processing of the reports without causing a loss of relevant information. This component is required to maintain both high recall and accuracy. For example, the phrases *as above, generally,* and *in particular* frequently occur in the reports but do not add relevant semantic information, and therefore can be ignored by the processor. Because the processor requires that the sentences must be parsed strictly according to the well-formed semantic structures and patterns specified by the semantic grammar, these phrases in a sentence would ordinarily cause the parser to fail. However, if they are explicitly listed as ignorable, these phrases can be safely skipped.

In order to translate the original sentence into a structured form, the processor uses translation rules that are defined along with the grammar structures. This is a compositional approach that is based on the work of Montague[38] and Gazdar et al.[39] In this approach, the translation of the structure is specified as a particular combination of the translations of the components. An example of the definitions of two grammar structures along with their corresponding translation rules are presented in Figure 5.

As shown in Figure 5, the translation rules occur after the structural specification of the grammar structures and are written after the colon (:). The rules may specify literal strings in addition to the translations of the components. The translation of a component is represented by enclosing the component in angle brackets and the inclusion of literals is represented by enclosing the literal strings in quotations. Thus the translation of **FindingPhr** consists of a target form containing the literal string **(Central)** → followed by the translation of the component **Findingterm**, the literal string **(Finding Mod)** → followed by the translation of **Lmods**, and the literal string **(Finding Mod)** → followed by the translation of **Rmods**. Since **Lmods** and **Rmods** are optional, they may not be present, in which case the corresponding **Finding Mod** relation is not added to the translation. The translation of **Findingterm** is specified in a similar manner, except that its components are not complex grammatical structures but correspond to atomic elements that are semantic categories. In this case, the translation of the element is directly associated with a word or phrase in the sentence, and is therefore specified by the target form in the lexical entry of the corresponding word or phrase.

As an example of how the processor utilizes the grammar shown in Figure 5, we describe the parse and translation of the sentence *mild cardiomegaly*. The translation is shown in Figure 6. The parse consists of a **FindingPhr** structure with an **Lmods** component and a **Findingterm** component. **Findingterm** has the value **cfinding**, which is an atomic element that corresponds to the word *cardiomegaly* in the sample sentence. According to the lexical entry of *cardiomegaly*, its semantic category is **cfinding** and its target form is **cardiomegaly**. The translation of **cfinding** is therefore **cardiomegaly** enclosed in square brackets. The

```
<FindingPhr>  --> {<Lmods>} <Findingterm> {<Rmods>} :
                   "(Central)->"<Findingterm>
                   "(Finding Mod)->"<Lmods>
                   "(Finding Mod)->"<Rmods>.

<Findingterm> --> disease : "["<disease>"]" |
                  cfinding : "["<cfinding>"]" |
                  pfinding : "["<pfinding>"]" |
                  descriptor : "[" <descriptor>"]".
```

Figure 5 Grammar definitions with translation rules.

square brackets are added as a result of the translation rule associated with **cfinding**. Thus, according to the translation rule of **FindingPhr**, the part of the translation of **FindingPhr** associated with the component **Findingterm** consists of the relation **Central** which has the value **[cardiomegaly]**. The rest of the translation of **FindingPhr** is associated with the translation of **Lmods**. **Lmods** has a degree modifier relation **(Degree Mod)** with the value **mild** enclosed in square brackets. In Figure 6 we eliminate the intermediate relation **Finding Mod** and show its value instead to simplify the figure.

Two more examples of the structured outputs that are produced by the processor using the semantic grammar are also shown in Figure 6. The original sentence is shown preceding the output form. The structured forms are consistent with the formal representational model of the domain although some of the names of the relations and concepts have been abbreviated and some relations consolidated to simplify the illustration. The concept **Rad Finding Structure**, which represents the structured form of a radiology finding, is shown in the figure as **Finding Str**, and the relation **Central Finding** appears as **Central**. The relation **Finding Mod** is not shown, but the individual modifier relations comprising the modifier, **Certainty Mod**, **Degree Mod**, **Change Mod**, and **Status Mod** are shown.

In Figure 6, the structured output for the second sentence *heart shows extensive enlargement* consists of a central finding relation **Central** which has the value **enlarged**. It also has a body location modifier **Bodyloc Mod** with the value **heart**, a degree modifier **Degree Mod** with the value **extensive**, and a certainty modifier **Certainty Mod** with the value **show**. These values do not yet represent controlled vocabulary concepts because they correspond to the initial output of the processing. These forms are shown again in Figures 8 and 10, which represent the output for subsequent stages of processing. The third example in Figure 6 shows the initial structured output for the sentence *decreasing but persistent right upper lobe infiltrate*. Its central finding is **infiltrate**, which has a body location modifier **right upper lobe**, a change modifier **decrease**, and a status modifier **persistent**.

```
Mild cardiomegaly.
[Finding Str]-
              (Central)->[cardiomegaly]
              (Degree Mod)->[mild].

Heart shows extensive enlargement.
[Finding Str]-
              (Central)->[enlarged]
              (Bodyloc Mod)->[heart]
              (Degree Mod)->[extensive]
              (Certainty Mod)->[show].

Decreasing but persistent right upper lobe infiltrate.
[Finding Str]-
              (Central)->[infiltrate]
              (Change Mod)->[decrease]
              (Bodyloc Mod)->[right upper lobe]
              (Status Mod)->[persistent].
```

Figure 6 Initial structured output forms.

The Phrase-regularization Component

After the clinical information from the reports has been structured, the next stage of processing consists of a compositional component that regularizes the output forms of phrases that are not contiguous. This is a critical step that further reduces the variety that occurs in natural language. For example, the phrase *hilar adenopathy* is contiguous and corresponds to the concept **hilar adenopathy** in our controlled vocabulary. When this phrase occurs in the reports, however, it appears in many variant forms, frequently with additional modifiers, such as *hilar and mediastinal adenopathy, adenopathy in the left hilus, significant adenopathy, left hilus,* and *adenopathy noted in left hilus*. Generally the variant forms are so diverse that it would be impossible to enumerate all of them for each multi-word phrase. Instead, a set of mappings representing their compositional structures are maintained. When a report sentence is processed, the initial structured form is compared with the compositional mappings to determine whether any variant forms of multi-word phrases occur in the sentence. If a variant form is found, it is replaced by its corresponding target form.

The mappings are maintained automatically. Because a multi-word phrase is a typical finding in radiologic reports, and therefore consists of a central concept and modifiers, it has a formal representation just like any other typical finding in a report. Its representation can be created automatically by treating the multi-word phrase like a sentence and by processing it to obtain its structured form. When the processor produces a structured form for the phrase, it is saved in a knowledge base of mappings representing the compositional structures of decomposable phrases. Figure 7 shows three examples of compositional mappings. The first mapping shows the compositional structure of *enlarged heart*. The first argument of the mapping, which is the value of the central finding, is used for indexing. The second argument of the mapping contains the structured form of the phrase and the third argument contains its target form. The mapping specifies that a structured form, consisting of a finding structure for which the central concept

```
mapping(enlarged,[[FindingStr]-
                    (Central)->[enlarged]
                    (Bodyloc Mod)->[heart]],
        'enlarged heart').

mapping(adenopathy,[[Finding Str]-
                     (Central)->[adenopathy]
                     (Bodyloc Mod)->[hilum]],
        'hilar adenopathy').

mapping(size,[[Finding Str]-
                    (Central)->[size]
                    (Change Mod)->[decrease]
                    (Bodyloc Mod)->[heart]],
        'decrease in heart size').
```

Figure 7 Compositional mappings.

```
Heart shows extensive enlargement.
[Finding Str]-
              (Central)->[enlarged heart]
              (Bodyloc Mod)->[heart]
              (Degree Mod)->[extensive]
              (Certainty Mod)->[show].
```

Figure 8 Structured output form after phrasal regularization.

is **enlarged** and the **bodyloc** modifier is **heart**, has a target form **enlarged heart**. A sentence such as *the heart appears to be severely enlarged* would, after being processed, have a finding structured form where the central concept is **enlarged**. It would also have a **bodyloc** modifier with the value **heart**, a **degree** modifier with a value **severe**, and a **certainty** modifier with a value **appears**. Because this form contains a central concept and bodyloc modifier that match the target form of the mapping for **enlarged heart**, the value of the central concept will be changed from **enlarged** to **enlarged heart**, but the modifier information will remain the same.

Figure 8 shows the structured output after the compositional matching is performed on the initial output of *heart shows extensive enlargement*, which is shown in Figure 6. The value of the central finding has been changed from **enlarged** to **enlarged heart** because the compositional mapping of **enlarged heart** was found to be subsumed by the initial structured form that was computed for the sample sentence. No change was made in the structured output of the second sentence in Figure 6 because it did not contain any phrases that were interrupted.

The mappings are created automatically by processing all phrasal lexical entries that begin with the symbol **phrase**. **Phrase** is used to specify that a phrase may occur in a non-contiguous variant form. For example, in Figure 4, *enlarged heart*, *hilar adenopathy*, and *decrease in size of heart* are all designated as phrases that may be non-contiguous. Atomic phrases, which are specified using the symbol **aphrase**, always occur contiguously and therefore mappings are not created for them.

The Encoder: Obtaining the Controlled Vocabulary

Mapping the regularized structured forms to controlled vocabulary concepts is the final stage of the processing. This is accomplished using a knowledge base specifying synonymous terms. The synonym knowledge base consists of associations between standard output forms and controlled vocabulary concepts. For example, this knowledge base contains an entry for the form **enlarged heart** associating it with the controlled vocabulary concept **cardiomegaly**. Figure 9 shows some examples of these entries. The first argument of the synonym specification is the target or standard form of the textual phrase, the second is the controlled vocabulary concept, and the third is the semantic category of the synonym. If a target form is identical to a controlled vocabulary concept, it does not have to be specified in this knowledge base. For example, the target form **cardiomegaly** associated with the word *cardiomegaly* directly corresponds to a controlled vocabulary concept and it is not specified as a target form in the synonym knowledge base to avoid redundancy.

In the encoding phase, if the structured output contains any values that match the first argument of a synonym entry in the synonym knowledge base, and if the semantic type of the form also matches, the second argument of the synonym entry, which is the controlled vocabulary concept, is substituted for the original value. At the end of this stage of processing, the only values that are in the structured form are unique controlled vocabulary concepts.

```
synonym('appear','moderate certainty',certainty).
synonym('calcified primary complex','Gohn complex',cfinding).
synonym('enlarged heart','cardiomegaly',cfinding).
synonym('nodular density','nodular opacity',pfinding).
synonym('severe','high degree',degree).
synonym('smaller heart','decrease in heart size',cfinding).
synonym('show','moderate certainty',certainty).
```

Figure 9 Synonym knowledge base.

```
Heart shows extensive enlargement.
[Finding Str]-
              (Central)->[cardiomegaly]
              (Bodyloc Mod)->[heart]
              (Degree Mod)->[high degree]
              (Certainty Mod)->[moderate certainty].
```

Figure 10 Final structured output form with controlled vocabulary concepts.

Figure 10 shows the final stage of the structured form for the sentence *heart shows extensive enlargement*. This form contains only concepts that are in the controlled vocabulary. This target structure was obtained by mapping the regularized forms shown in Figure 8 to the controlled vocabulary concepts. By comparing the two figures, we can see that the finding **enlarged heart** was mapped to **cardiomegaly**, the degree modifier **extensive** was mapped to **high degree**, and the certainty modifier **show** was mapped to **moderate certainty**. This final mapping was facilitated by the first two stages of processing, which structured and regularized the contiguous and non-contiguous expressions in the original sentence. If the compositional stage of processing were missing, it would be impossible to enumerate all variations of **enlarged heart**. Therefore, certain variations would be missed and consequently would not be mapped to the controlled vocabulary concept **cardiomegaly**.

Immediately before this final stage of the processing, the reports were structured and variant forms regularized. The variety of expression that is intrinsic to natural language was already drastically reduced, facilitating the final mapping to controlled vocabulary concepts. The synonym knowledge base is simple to specify and utilize only because all the variants forms have already been mapped to standard phrases. If the variant forms were not mapped to standard phrases, this knowledge base and the associated mappings would be much more complex.

After the encoded form is obtained, a post-processing clean-up procedure is performed to resolve inconsistencies that occur when a sentence contains both negation and certainty information, such as *evidence of enlarged heart not observed*. The target structure of negation information is a qualifier **Certainty Mod** whose value is **no**. The qualifier **Certainty Mod** is also a target structure for certainty information, such as *possibly*, except its value does not denote negation. Thus, the output for the sentence *evidence of edema not noted* would consist of a central finding whose value is **edema**, a certainty modifier whose value is **high certainty**, corresponding to *evidence of*, a certainty modifier whose value is **no**, corresponding to *not*, and a certainty modifier whose value is **high certainty** corresponding to *noted*. It is logically inconsistent to associate these different certainty values with a finding.

The inconsistency is resolved by assigning an order of precedence to each value ranging from 5 for **no** to 0 for **high certainty**. These values represent the five different values associated with certainty in our model. Whenever a target structure corresponding to a finding with qualifiers contains more than one certainty value, the one with the highest precedence is chosen. Thus, for a sentence containing negation and other certainty information, as in *evidence of edema not noted* and *no edema noted*, it will be determined that the value for **Certainty Mod** should be **no**. Another adjustment is also made for sentences containing *no . . . or . . .*, such as *no edema or pleural effusion*, so that the final interpretation consists of two findings *no edema noted* and *no pleural effusion noted*.

Results

A preliminary study consisting of 230 reports that were randomly selected was undertaken to evaluate the processor. Four diseases: neoplasm, congestive heart failure, acute bacterial pneumonia, and chronic obstructive pulmonary disease, were chosen for the evaluation. The selection of the diseases was done by an independent physician based on McDonald's work.[40] Because we wanted to determine how the text-processing system performed when it was not trained specifically for the four diseases, no change was made to the system once the four diseases were selected. Thus no new words or phrases associated with the diseases were added to the lexicon, and no new controlled vocabulary terms were added to the MED.

Three physicians were given a listing in which each report was followed by a checklist of the four diseases. The physicians manually read each report and checked off the disease(s) that they felt were likely to be present according to the information in the report. The variance among the three physicians was great, and we concluded that more than three physicians were needed for a definitive study. In addition, more reports were also needed because certain conditions, such as chronic obstructive pulmonary disease, did not occur frequently enough to permit statistically significant results. We feel a subsequent evaluation should be performed utilizing more physicians and more reports.

Automated queries were also written to retrieve reports associated with the four conditions. The automated queries utilized the structured outputs obtained from the text processor. A comparison was made between the results obtained manually and those obtained automatically. The "gold standard" for the automated system consisted of those reports checked by at least two of the three physicians. Recall and precision of the automated system for the four conditions were 70% and 87%, respectively. The recall number was lower than anticipated because some of the retrievals were not sufficiently trained. The

processor itself performed very well, because a larger percentage of outputs were structured properly but were not retrieved by the queries. When the queries were corrected, the recall of the system improved considerably: recall was 85% and precision remained the same.

We anticipate that the recall of the processor will be improved when more words and phrases are added to the lexicon. Presently, words and phrases that are clinically relevant but that occur relatively infrequently are being added in the lexicon. However, it is possible that extending the lexicon to increase the recall could also have a negative effect on the precision of the system. Spelling errors also lowered the recall of the automated system. We do not offer a spelling-correction feature because it will shortly become available at CPMC with the introduction of a commercial word-processing system.

Another significant result of the evaluation showed that certain ways of organizing the controlled vocabulary are more conducive to accurate retrieval. Although the proper structuring of the text is a critical task in representing the salient information, it is not the only task associated with the structured forms. Automated queries subsequently utilize the structured forms to retrieve findings. In some circumstances the findings in the text were structured properly, but writing the query based on the encoded terms in the structured output form was not straightforward. The difficulties could be corrected by modifying the controlled vocabulary. Our results showed that multi-word terms, such as **pleural effusion**, are more likely to be included in a query than a singular term, such as **effusion**, with specific modifiers. We plan to modify our controlled vocabulary accordingly.

Our results also showed that some queries are inherently more difficult to write than others. The query to retrieve reports for neoplasm was more complex than the query to retrieve reports for congestive heart failure. The phrases implying congestive heart failure were fewer and less diverse, and therefore, both the processor and the query were effective without performing training just for congestive heart failure. There were many more words and phrases associated with neoplasm than with congestive heart failure. Therefore, to maximize effectiveness, it would have been beneficial to train both the natural-language processor and the automated retrieval for neoplasm. For the processor, training would generally mean extending the lexicon to add more words and phrases associated with neoplasm. For example, **metastases** and **metastatic disease** are in the lexicon currently, but **metastatic deposits** and **bump**, which were found in two new reports, are not. For the query, training would mean adding more controlled terms to the query.

Discussion

We chose the impression section of chest x-rays for our initial application because they encompass a complex domain of clinical medicine, contain a limited but substantial variety of body sites and internal problems, and yield useful clinical information for decision support and research. In addition, online radiology reports are readily accessible from the central patient database at CPMC. Although the processor was evaluated using a random sample of 230 reports, the processor has been used to codify the impression section of 8,000 chest x-ray reports.

The text processor, using only the semantic grammar, was able to process the bulk of the clinical information in the domain of impressions of chest x-ray reports. As described above, semantic techniques work very well for simple sentences that contain common expressions or semantic patterns in the domain. The semantic component was effective for the given domain because the reports are generally simple and naturally well structured. We believe this technique will also be successful for a substantial portion of the text sentences in many other domains (e.g., pathology impressions, echocardiogram findings, endoscopy findings, orders). However, the results would probably be not as good for less structured text (e.g., discharge summaries, admission and surgical notes, progress reports) unless a syntactic component were also used, and therefore we believe that a syntactic component is critical if the system is to be successfully extended to those domains. The advantages of the semantic grammar is that the processing is accurate and efficient, and the semantic grammar can be developed relatively quickly. It took approximately half a person year to develop it.

The system will be extended to handle a broader variety of text by adding a syntactic grammar, which is currently under development. A medium-sized syntactic grammar has been written and is being tested. The syntactic grammar is much more complex and therefore requires a longer time to develop. We have been working on it for approximately two person years.

In discussing the text-processing system, we also briefly discussed the design of a formal model that represents the informational content of the reports. The most important design criterion for our model

involved the balancing of conflicting requirements: the need for a complete representation of the information in the reports against the need for a representation that is not overly complex for use by decision support or other computerized processes. This entails that the informational content of the narrative be represented in a simple form while covering all types of critical information. For example, adding information expressing fuzziness and uncertainty increases the complexity of a system very much but is essential to represent. This type of information, which occurs frequently with radiology findings, as in *mild opacity in left lower lobe, slight decrease in opacity in left lower lobe,* and *no opacity in left lower lobe,* significantly changes the meaning of the findings and thus must also be represented. However, this also means that retrieval of findings is more complicated because different modifiers must also be checked as they critically affect the underlying meaning.

Another crucial knowledge-representation issue that we discussed in this paper concerns the type of additional knowledge that is required to facilitate the mapping of natural-language expressions into unique medical concepts. Although the parsing component of the language processor significantly reduces the variety of ways of expressing information to standardized linguistic forms (e.g., the standard form for *cardiac enlargement, enlargement of heart,* and *heart is enlarged* is **enlarged heart**), for each standard form there is usually a set of completely different linguistic forms that correspond to the same concept. In order for the retrieval of clinical information to be reliable, all the forms in the set have to be associated with the same concept. This is done using a synonym knowledge base that associates standard forms of words and phrases with concepts in the controlled vocabulary. For example, the form **enlarged heart** is associated with the concept **cardiomegaly**, which is in the controlled vocabulary.

Conclusions

We have developed and evaluated a text processor that extracts and structures clinical information from textual radiology reports and translates the information to terms in a controlled vocabulary so that the clinical information can be accessed by further automated procedures. Although the processor has been applied initially to the domain of chest x-ray impressions, the methodology is principled, modular, and extensible, and, we believe, can readily be ported to other clinical domains.

Our approach combines advantageous elements of different methodologies (pattern matching, semantic-based, and syntactic-based) into one uniform framework where the text processing algorithm is always the same, and only the grammars contain different language capabilities. This approach maximizes accuracy and efficiency because a semantic grammar is used initially. If the text cannot be handled using a semantic grammar, analysis will be attempted using a syntactic grammar. We have not yet added the syntactic grammar because it is still being tested. However, for the radiology domain, the semantic grammar is quite effective. This type of system can be developed rather quickly, because initially it can be built to handle limited information, and then it can be extended incrementally to handle a broader range of information without changing the underlying approach.

The potential for this methodology is to make available a large body of clinical information that would otherwise be inaccessible for applications other than manual physician review. This methodology is not intended to replace coded data entry, but does offer an alternative when coded entry is not practical or acceptable to health care providers. Another potential for this processor is that reports from any radiology site and from any previous time period can be processed if electronic versions of the reports are made available. The result would be virtual enrollment of previously evaluated patients, which would greatly enhance the value of longitudinal and outcome studies.

The authors gratefully acknowledge Dr. Paul Clayton for providing invaluable support along with the resources of the Center for Medical Informatics at Columbia Presbyterian Medical Center. They are also very grateful to Dr. George Hripcsak for providing insights concerning the development of the controlled vocabulary and for his design and analysis of the evaluation.

References ■

1. Benoit RG, Cushing BM, Teitelbaum SD, Van Wijngaarden MH. Direct physician entry of injury information and automated coding via a graphical user interface. In: Frisse ME, ed. Proc Sixteenth Annu Symp Computer Applications in Medical Care. New York: McGraw-Hill, 1992:787–8.
2. Bell DS, Greenes RA, Doubilet P. Form-based clinical input from a structured vocabulary: initial application in ultrasound reporting. In: Frisse ME, ed. Proc Sixteenth Annu Symp Computer Applications in Medical Care. New York: McGraw-Hill, 1992:789–90.
3. Cristea D, Mihaescu T. Combining menus with natural language processing in recording medical data. Clin Comput. 1988;16(5):156–166.
4. Lin R, Lenert L, Middleton B, Shiffman S. A free-text processing system to capture physical findings: canonical phrase identification system (CAPIS). In Clayton PD, ed. Proc Fifteenth Annu Symp Computer Applications in Medical Care. New York: McGraw-Hill, 1991:843–7.

5. Johnson K, Poon A, Shiffman S, Lin R, Fagan L. A history-taking system that uses continuous speech. In: Frisse ME, ed. Proc Sixteenth Annu Symp Computer Applications in Medical Care. New York: McGraw-Hill, 1992:757–61.
6. Shiffman S, Lane C, Johnson K, Fagan L. The integration of a continuous-speech-recognition system with the QMR diagnostic program. In: Frisse ME, ed. Proc Sixteenth Ann Symp Computer Applications in Medical Care. New York: McGraw-Hill, 1992:767–71.
7. Linn NA, Rubenstein RM, Bowler AE, Dixon JL. Improving the quality of emergency department documentation using the voice-activated word processor: interim results. In: Frisse ME, ed. Proc Sixteenth Annu Symp Computer Applications in Medical Care. New York: McGraw-Hill, 1992:772–6.
8. Voice Med. Waltham, MA: Kurzweil Applied Intelligence Inc., 1991.
9. Rothwell D, Hause L, Frey C. Lab management memo. Chicago: College of American Pathology, May 1982.
10. Gabrieli E, Speth D. Computer processing of discharge summaries. In: Stead WW, ed. Proc Eleventh Annu Symp Computer Applications in Medical Care. Los Angeles: IEEE Computer Science Press, 1987:137–40.
11. Gell G. Free text processing in clinical documentation. Clin Comput. 1982;10:170–9.
12. Grams R, Jin Z. The natural language processing of medical databases. Med Syst. 1989;13:79–87.
13. Hersh WR, Greenes RA. SAPHIRE: an information retrieval system environment featuring concept-matching, automatic indexing, and probabilistic retrieval. Comput Biomed Res. 1990;23:405–20.
14. Lin R, Lenert L, Middleton B, Shiffman S. A free-text processing system to capture physical findings: canonical phrase identification system (CAPIS). In: Clayton PD, ed. Proc Fifteenth Annu Symp Computer Applications in Medical Care. New York: McGraw-Hill, 1991:168–72.
15. Zingmond D, Lenert LA. Monitoring free-text data using medical language processing. Comput Biomed Res. 1993;26:467–81.
16. Baud RH, Rassinoux AM, Scherrer JR. Knowledge representation of discharge summaries. In: Stefanelli M, Hasman A, Fieschi M, Talmon J, eds. Proc Third Conf Artificial Intelligence Med, Maastricht, The Netherlands. Berlin: Springer-Verlag, 1991:173–82.
17. Baud RH, Rassinoux AM, Scherrer JR. Natural language processing and semantical representation of medical texts. Meth Inform Med. 1992;31:117–25.
18. Rossi-Mori A. CEN/TC251/PT003 model for representation of terminologies and coding systems in medicine. In: Proc Opportunities for European and US Cooperation in Standardization in Health Care Informatics [Seminar]: Geneva, Switzerland, 1992.
19. Canfield K, Bray B, Huff SM. Representation and database design for clinical information. In: Miller RA, ed. Proc Fourteenth Annu Symp Computer Applications in Medical Care. Los Alamitos, CA: IEEE Computer Society Press, 1990:350–3.
20. Ranum DL, Haug PG. Knowledge based understanding of radiology text. In: Greenes RA, ed. Proc Twelfth Annu Symp Computer Applications in Medical Care. Washington, DC: IEEE Computer Science Press, 1988:141–5.
21. Minsky M. A framework for representing knowledge. In: Haugland J, ed. Mind Design. Cambridge, MA: MIT Press, 1981:95–128.
22. Sowa JF. Conceptual Structures. Reading, MA: Addison-Wesley, 1984.
23. Canfield K, Bray B, Huff SM, Warner H. Database capture of natural language echocardiology reports: a UMLS approach. In: Kingsland LC, ed. Proc Thirteenth Annu Symp Computer Applications in Medical Care. Washington, DC: IEEE Computer Society Press, 1990:559–63.
24. Sager N, Friedman C, Lyman MS, et al. Medical language processing: computer management of narrative data. Reading, MA: Addison-Wesley, 1987.
25. Nhan NT, Sager N, Lyman M, Tick LJ, Borst F, Su Y. A medical language processor for two Indo-European languages. In: Kingsland LC, ed. Proc Thirteenth Annu Symp Computer Applications in Medical Care. Washington, DC: IEEE Computer Society Press, 1989:554–8.
26. Sager N, Lyman M, Nhan NT, Tick LJ, Borst F, Scherrer JR. Clinical knowledge bases from natural language patient documents. In: Lun KC, Degoulet P, Plemme TE, Rienhoff O, eds. MEDINFO 92. Amsterdam, The Netherlands: North-Holland, 1992:1375–81.
27. 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;1:35–50.
28. Friedman C, Hripcsak G, Johnson SB, Cimino JJ, Clayton PD. A generalized relational schema for an integrated clinical patient database. In: Miller RA, ed. Proc Fourteenth Annu Symp Computer Applications in Medical Care. Los Alamitos, CA: IEEE Computer Society Press, 1990:335–9.
29. Johnson SB, Friedman C, Cimino JJ, Clark AS, Hripcsak G, Clayton PD. A conceptual schema for a central patient database. In: Clayton PD, ed. Proc Fifteenth Annu Symp Computer Applications in Medical Care. New York: McGraw-Hill, 1991:381–7.
30. Evans D, Chute C, Cimino J, et al. CANON: towards a medical concept representation language for electronic medical records (abstr). In: Kahn MG, ed. Proc 1993 Spring Congress of the American Medical Informatics Association. Bethesda, MD: American Medical Informatics Association, 1993:26.
31. 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.
32. Campbell KE, Musen MA. Representation of clinical data using SNOMED III and conceptual graphs. In: Frisse ME, ed. Proc Sixteenth Annu Symp Computer Applications in Medical Care. New York: McGraw-Hill, 1992:354–8.
33. Campbell KE, Musen MA. Creation of a systematic domain for medical care: the need for a comprehensive patient-description vocabulary. In: Lun KC, Degoulet P, Plemme TE, Rienhoff O, eds. MEDINFO 92. Amsterdam, The Netherlands: North-Holland, 1992:1437–42.
34. Rector AL, Nowlan WA, Kay S. Conceptual knowledge: the core of medical information systems. In: Lun KC, Degoulet P, Plemme TE, Rienhoff O, eds. MEDINFO 92. Amsterdam, The Netherlands: North-Holland, 1992:1420–6.
35. Friedman C, Cimino JJ, Johnson SB. A conceptual model for clinical radiology reports. In: Safran C, ed. Proc Seventeenth Annu Symp Computer Applications in Medical Care. New York: McGraw-Hill, 1993:829–33.
36. Sebesta R. Concepts of Programming Languages. Redwood City, CA: Benjamin/Cummings, 1993.
37. Gazdar G, Mellish C. Natural Language Processing in PROLOG. New York: Addison-Wesley, 1989.
38. Montague R. Formal Philosophy. New Haven, CT: Yale University Press, 1974.
39. Gazdar G, Klein E, Pullum G, Sag I. Generalized Phrase Structure Grammar. Oxford, UK: Basil Blackwell, 1985.
40. McDonald CJ. Action-Oriented Decisions in Ambulatory Medicine. Chicago: Year Book Medical Publishers, 1981.