NATURAL LANGUAGE PROCESSING AND THE REPRESENTATION OF CLINICAL DATA

NAOMI SAGER, PHD, MARGARET LYMAN, MD,
CHRISTINE BUCKNALL, MD, NGO NHAN, PHD, LEO J. TICK, PHD

Reprinted from **JOURNAL OF THE AMERICAN MEDICAL INFORMATICS ASSOCIATION**
Published by HANLEY & BELFUS, INC., Philadelphia, PA

Research Paper ■

Natural Language Processing and the Representation of Clinical Data

NAOMI SAGER, PHD, MARGARET LYMAN, MD, CHRISTINE BUCKNALL, MD, NGO NHAN, PHD, LEO J. TICK, PHD

Abstract **Objective**: Develop a representation of clinical observations and actions and a method of processing free-text patient documents to facilitate applications such as quality assurance.

Design: The Linguistic String Project (LSP) system of New York University utilizes syntactic analysis, augmented by a sublanguage grammar and an information structure that are specific to the clinical narrative, to map free-text documents into a database for querying.

Measurements: Information precision (I-P) and information recall (I-R) were measured for queries for the presence of 13 asthma-health-care quality assurance criteria in a database generated from 59 discharge letters.

Results: I-P, using counts of major errors only, was 95.7% for the 28-letter training set and 98.6% for the 31-letter test set. I-R, using counts of major omissions only, was 93.9% for the training set and 92.5% for the test set.

■ **J Am Med Informatics Assoc.** 1994;1:142–160.

The attention currently being given to the computer-based patient record[1,2] adds impetus to the development of representational structures for clinical data. It is hoped that standardized structures can be developed to serve as a framework for combining patient data from multiple sources. In the evolution of computerized patient record systems, the controversy between free text and preset categories for recording patient data has not been resolved. The need for standards pushes toward preset categories and controlled vocabularies, while the need for expressive power, so as not to distort the patient data, speaks for allowing some amount of free-text reporting. A compromise that is not compromising is called for. It is the aim of this paper to show that the techniques of linguistic analysis and natural-language processing (NLP) can contribute to this effort.

Within medical informatics there has been a long-standing concern with medical language. In a 1973 review,[3] Pratt (then chief of the Computer Research Branch at the National Institutes of Health) emphasized that the data underlying the patient care process "are in the large majority nonnumeric in form and are formulated almost exclusively within the constructs of natural language. . . The data are language data." These constructs were identified as both syntactic and semantic, and were important in the development of the multifaceted Systematized Nomenclature of Pathology (SNOP),[4] later SNOMED,[5] now SNOMED International (SNOMED III).[6] The possibility of automatically encoding pathology diagnostic statements into the SNOP code was successfully demonstrated by the operation of the encoder developed at the NIH.[7] Work has continued on automated indexing from natural-language clinical documents into SNOMED and International Classification of Diseases (ICD) codes by researchers in Europe, Canada, and the United States, in part collaboratively.[8,9]

Interest in unifying the medical vocabularies of different medical knowledge sources in order to facilitate the use of knowledge resources across information systems led to the Unified Medical Language System (UMLS) initiative of the National Library of Medicine.[10] The UMLS Metathesaurus is intended to be a lexical framework for this integration.[11] Exten-

Affiliation of the authors: New York University, New York, NY.

Correspondence and reprints: Naomi Sager, PhD, Courant Institute of Mathematical Sciences, New York University, 251 Mercer Street, New York, NY 10012.

Received for publication: 6/24/93; accepted for publication: 10/04/93.

sions to the UMLS Metathesaurus for more complete coverage of clinical vocabulary are under way, for example, to include laboratory terminology.[12] In addition, there are experiments in applying statistical procedures to massive text files in order to establish semantic relationships among natural-language expressions in free-text documents and in established vocabularies,[13–15] and there exists a system of deeply coded word strings to match against free-text patient documents.[16, 17]

It is not easy to separate the activity of classifying the words and phrases characteristic of a discipline (its free-text vocabulary) from a consideration of the structures that the words and phrases are part of, or from the tasks to which the resultant structured vocabulary is to be applied. The issues are yet more complex when instead of speaking of "information," seen as more closely tied to factual reports of named observables, the topic becomes "knowledge" with its penumbra of generalizations, models, and inference procedures, related to human intellectual tasks that draw upon all of these. Medical informatics, from early in its evolution, was concerned with these issues[18,19] and contributed pioneering systems of knowledge-based medical decision support.[20–23] The development of important clinical knowledge bases such as QMR,[24] DXplain,[25] and the knowledge base in the HELP system[26] has posed the question of the sharing and "re-use" of these resources[27] with, as one approach, the adoption of a common "interlingua," using frames as the formalism.[28]

If we restrict our discussion to clinical information (not medical knowledge in the broader sense), that is, to the facts about patients and the care given them, such as would reside in a database to be used for quality assessment or for clinical research, then the problem of finding a representational framework is perhaps tractable. One approach is to base the representation on the categories and structures used explicitly and implicitly in the documents that physicians have traditionally used to report these facts. The language of clinical reporting is stylized with regard to both the division of content into document subsections and the types of sentences to be found in each subsection. Aside from the few sentences or subsections that describe the medical reasoning behind decisions, the connections among sentences are for the most part either of the "*and*" type, i.e., one finding after another, as in the physical examination (objective) subsections, or of the implicit or stated "*and then*" type, as in the history (subjective) and course subsections of many hospital discharge summaries. A number of other connectives (e.g., *with, secondary to*) play significant roles.

What is described here is a representation of clinical observations and actions and a method of processing free-text patient documents that make use of the inherent regularities in the language of clinical reporting. Documents are analyzed sentence by sentence into instances of the representation and mapped into a database for querying. By using a database as the final repository we make possible a test of the efficacy of the representation by incorporating the processor into an application (quality assurance) whose results can be evaluated by trained medical reviewers.

A Theoretical Basis for Language Processing

In the 1960s the National Science Foundation and other agencies sponsored basic research in language and information as the basis for information systems of the future. It was anticipated at that time that automated language analysis would provide the bridge between users and the stored knowledge they needed. The first step was understood to be a parser for a broad segment of the English language. The Linguistic String Project (LSP) at New York University produced such a parser, along with a programming language for natural-language grammars and a comprehensive computer grammar of English.[29–32]

A necessary supplement to syntactic analysis, in order to arrive at a representation of the specific information in a text or query, is a method of determining the semantic categories of the discourse. The linguist Z. Harris provided a basis for developing the relevant categories and relations in technical and scientific subject matters when he posited the existence of "sublanguage grammars,"[33] later developed more fully.[34,35] Briefly, the additional constraints (over and above the grammatical rules of a language) on what constitutes an acceptable sentence within an established discipline can be formulated in a sublanguage grammar for that discipline (e.g. *Ions enter the cell* but not *The cell enters ions*; rules that exclude the latter would be part of the sublanguage grammar of cell physiology). The sublanguage method has proved productive in the design and development of a number of natural-language information systems.[36–38]

Subsequent work, in particular on medical sublanguages,[39–41] has shown that data structures ("information formats") corresponding to the main statement types of a sublanguage can be defined.[42,43] The statement types are based on recurring patterns of sublanguage (here medical) word-class co-occurrence in syntactic relations (e.g., a symptom word with a body-part word or phrase as modifier: *epigastric pain, pain in right lower quadrant*). While these patterns do not reflect nuances of usage, which may be very

important in some contexts, they do reflect what is most usual and hence most relevant for review and comparison with information from other sources. The information-formatting program of the LSP, which operates with the rules of a sublanguage grammar, i.e., with regularly occurring categories in their syntactic combinations, misses some details but on the whole converts the free-text information into a structured form with reasonable and consistent results.[44]

Medical areas of study and/or computer applications using the LSP system have included radiology (oncology follow-up, nuclear medicine reports),[45] pharmacology (literature),[42] sickle-cell disease,[46] pneumonia, bacterial meningitis,[47] anatomic pathology,[48] rheumatoid arthritis,[49] digestive surgery,[50] and asthma. Experience with this variety of medical language material has shown us that the principles of sublanguage analysis apply. Each medical discipline or subdiscipline expresses its content in relatively stereotyped sentence types based on its specialized word usage. We found that clinical summaries (admission histories, discharge summaries, reports of clinic visits, letters to referring physicians) were constructed in large part of the same gross stereotypical sentence types. As a result, the medical sublanguage grammar and the medical word classification scheme (Fig. 3, below), have remained stable over a range of clinical areas, and the lexicon that is coded with respect to these classes has required little modification except additions in moving from one disease area to another. If a detailed analysis is desired for the reports of particular tests and procedures (EKG, EEG, etc.), then further refinement of the statement types and of the word classes appearing in these statement types will be needed. This overall uniformity of linguistic structure down to the level of semantic relevance is what in effect makes medical language processing possible.

Representation of Clinical Narrative in the LSP System

The Basic Data Structure

As a data structure an Information Format (I-F) is a template for holding the words of a sentence (or sentence-part) that corresponds to a statement type of the sublanguage. Figure 1 shows a clinical sentence analyzed into I-F occurrences of the PATIENT STATE type, the most commonly occurring one in clinical documents. The slots of the template are in capital letters and the words of the sentence are in bold italics, followed by the medical lexical classes of the word(s) in the slot. Only the instantiated slots of the template appear in Figure 1; those not filled by sentence words are absent. The complete PATIENT STATE I-F, labeled I-F5, is shown in Figure 2. The I-Fs are much the same as those previously described,[44] as are the medical lexical classes, which are listed and illustrated in Figure 3.

In Figure 1, Part A, the tree-like structure of an I-F and of the connective structure relating instantiated I-Fs in a sentence is indicated by indentation. The sentence in Figure 1, *"This 29 year old girl who is known to be asthmatic for 15 years presented with a 4 day history of steadily increasing exertional wheezy dyspnoea with a cough productive of green sputum"* contains three instances of the PATIENT STATE I-F (I-F5) in the connective relation illustrated diagrammatically in Part B of Figure 1, where only the contents of the SIGN-SYMPTOM (S-S) slot and the DIAGNOSIS (DIAG) slot of each I-F5 are shown.

A difference between the instantiated PATIENT STATE I-F (Fig. 1) and the uninstantiated PATIENT STATE I-F (Fig. 2) is the presence in Figure 1 of the TIME unit, not shown in Figure 2. Modifiers of the types NEGATION, MODAL (uncertainty) and TIME, which can apply to any statement, are defined as separate modifier units and inserted into the instantiated I-F to which they apply before loading to a relational database. In the case where the connective contains a negative or uncertainty word (or word-part) that applies to its argument (e.g. *"without"*), a procedure removes the negative or uncertainty portion from the connective and creates a NEGATION or MODAL modifier node in the I-F of the argument (e.g., in effect, *"without complication"* → *"with no complication"*). The same holds true in the (rare) case where the embedding phrase of an EMBEDDED-OBJ contains the negative or uncertainty-marker (e.g., in effect, *"It is suspected that the patient is a substance abuser"* → *"It is that the patient is a suspected substance abuser"*).

On the whole there is almost a 1:1 correspondence between the main medical lexical classes (Fig. 3) and the similarly named slots of the I-F. In this case the syntactic procedures serve mainly to check on the validity of the segmentation into informational units and their expansion (around conjunctions) to form complete units, to correctly assign negation and uncertainty markers, and to identify time expressions for further processing.[51, 52]

Some slots, however, require in addition special computations involving both the syntax and the medical lexical attributes of text expressions. These computations can be packaged into macros associated with the I-F slot. An example is the INFLUENCE slot of the PATIENT STATE I-F. It is intended to capture expressions that describe conditions affecting the manifes-

tation of symptoms. Several types are recognized. The main ones are composed as follows:

1. Certain (mainly time) prepositions or subordinate conjunctions + expressions of the H-PTFUNC type (normal physiologic function), forming a modifier of H-INDIC (symptom): *epigastric pain **after eating**; chest pain **on exertion**; tightness in the chest **at rest**;*
2. A change word H-CHANGE having subattribute MORE or LESS + expressions of the H-PTFUNC type, forming a modifier of H-INDIC: *chest pain **increasing with exercise**; nausea **decreased by lying down**;*
3. A response word H-RESP + expressions of the H-PTFUNC type, forming a modifier of H-INDIC: *pain **relieved by resting**; cramps **eased by exercise**.*

Notice with regard to type 3 that *pain relieved by nitroglycerine* would be treated differently from *pain relieved by resting*, because *nitroglycerine* as a medication word H-TTMED requires a treatment I-F; this occurrence is represented as a compound treatment-response I-F via a relation (see Fig. 4. below).

The Larger Schema

Figure 4 shows schematically, in a form based on the entity-relationship (E-R) data model,[53] the main types of medical facts (in linguistic terms: medical statement types) that occur in clinical documents, and their associated medical lexical ("atomic") attributes. The medical lexical attributes of the data model in Figure 4 are the same as the medical lexical classes

A. COMPUTER OUTPUT FOR HISTORY SENTENCE

```
*SID=GLCBA 002B.1.01
* THIS 29 YEAR OLD GIRL WHO IS KNOWN TO BE ASTHMATIC FOR 15 YEARS PRESENTED
* WITH A 4 DAY HISTORY OF STEADILY INCREASING EXERTIONAL WHEEZY DYSPNOEA
* WITH A COUGH PRODUCTIVE OF GREEN SPUTUM .

(CONNECTIVE (REL-CLAUSE (CONN = 'who'))
      (CONNECTIVE (RELATION (CONN = 'with' (H-CONN)))
            (I-F5: PATIENT STATE
                  (PSTATE-SUBJ (PT = this 29 (QNUMBER) year (NUNIT NTIME1)
                                    old (H-AGE) girl (H-PT)))
                  (VERB = presented (H-PTVERB) with)
                  (TIME (TM-PERIOD = a history (H-TMPER))
                        (Q-N   (NUM = 4 (QNUMBER))
                               (UNIT = day (NUNIT NTIME1))))
                  (TENSE = [PAST])
                  (PSTATE-DATA (S-S = of exertional (H-PTFUNC) wheezy (H-INDIC)
                                    dyspnoea (H-INDIC))
                               (QUANT = steadily
                                    increasing (H-CHANGE [(MORE)])))))
            (I-F5: PATIENT STATE
                  (PSTATE-DATA (S-S = a cough (H-INDIC) productive
                                    of ('OF') green sputum (H-INDIC))))
      (CONNECTIVE (EMBEDDED (CONN = '[EMBEDDED-OBJ]'))
            (I-F00: SENTENTIAL OPERATOR
                  (VERB = is (VBE) known)
                  (TENSE = [PRESENT]))
            (I-F5: PATIENT STATE
                  (PSTATE-SUBJ (PT = girl (H-PT)))
                  (VERB = be (VBE))
                  (PSTATE-DATA (DIAG = asthmatic (H-DIAG)))
                  (TIME (TPREP1 = for ('FOR'))
                        (Q-N   (NUM = 15 (QNUMBER))
                               (UNIT = years (NTIME1)))))))
```

B. CONNECTIVE STRUCTURE DIAGRAMMATICALLY

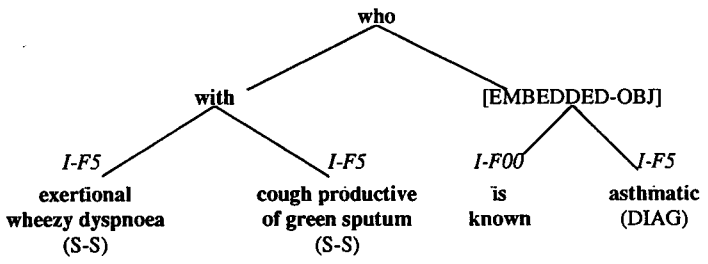

Figure 1 Information formatting output of the LSP language processor as applied to a representative sentence from the history paragraph of a patient document. Part A shows occurrences of the instantiated Information-Format of the PATIENT STATE type. Part B shows the linguistic connective structure of these occurrences in the sentence.

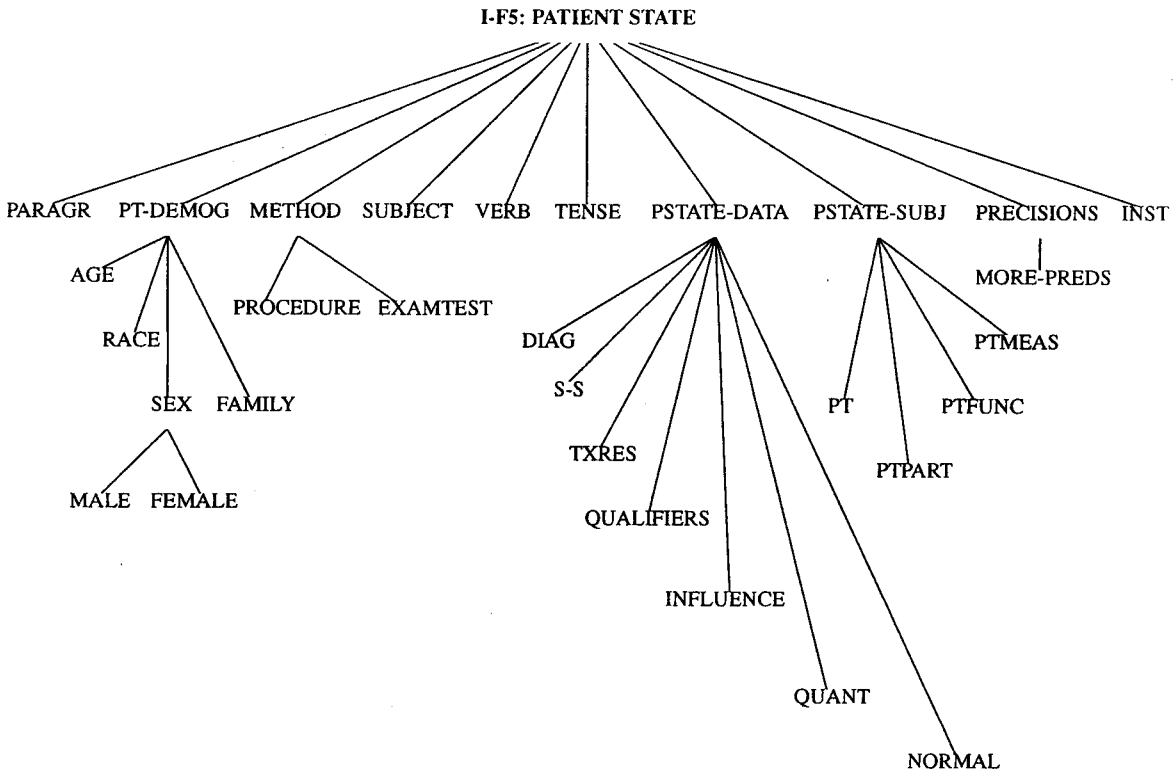

Figure 2 The complete uninstantiated PATIENT STATE Information-Format (I-F5). PSTATE-SUBJ holds the underlying medical subject of the reported observation. PSTATE-DATA contains the predicate (finding) about the medical subject. Nonmedical subjects and verbs default to SUBJECT and VERB. METHOD contains the procedure or physical examination test (if stated) that gives rise to the finding. PRECISIONS holds additional medical modifiers and INST holds mentions of institutional personnel or departments (not proper names).

(Fig. 3) used by the LSP system to arrive at the analysis of clinical documents.

The E-R model diagram does not fully display the internal structure of I-Fs, which in the E-R model are represented as FACT types. What the E-R diagram shows is that the statement of a medical fact (MEDICAL FACT box in Fig. 4), whether of the CLINICAL, LABORATORY, TREATMENT, or RESPONSE subtype, is composed of a subject and a predicate (SUBJECT and PREDICATE boxes), each of which has associated with it a set of atomic attributes, which are listed in attached boxes. The SUBJECT may be physically absent in the statement being modeled, but if so, it is implicit.

In the E-R diagram, the MEDICAL FACT as a whole carries the evidential attributes (H-NEG, H-MODAL) and all the temporal attributes (H-TMXXXX), seen in a box at the lower left attached to the MEDICAL FACT box. This reflects the fact that evidential and temporal modifiers can occur on any type of clinical statement. In an occurrence of a CLINICAL FACT type, a PROCEDURE may be mentioned as the source of the finding; this may be a physical examination procedure H-TXCLIN, e.g., *"auscultation,"* or other procedure H-TXPROC, e.g., *"x-ray,"* as indicated by the upper left PROCEDURE box with its associated atomic attributes.

The report of a laboratory fact (LABORATORY FACT box) is characterized by the description of a test (TEST box) with its associated attributes (H-TXSPEC, H-TXVAR, H-PTSPEC, H-ORG) shown in the box attached to the TEST box. These attributes are the names of the lexical classes that distinguish the laboratory statement type from others. The TREATMENT FACT type is likewise distinguished from other statement types by the lexical classes ("atomic attributes") shown in the box attached to the TREATMENT FACT box, and similarly for the RESPONSE FACT type.

While the four subtypes (CLINICAL, LABORATORY, TREATMENT, RESPONSE) of MEDICAL FACT seen in the E-R diagram correspond to the main I-Fs of the LSP language processor, the further subtyping in the E-R diagram has a different significance. The CLINICAL FACT type is subtyped according to the paragraph structure of the documents being processed; I-F occurrences are labeled in the database as to which

paragraph they occurred in. Shown in Figure 4, to illustrate, are subtypes corresponding to several named paragraphs frequently found in discharge summaries: DIAGnosis, HISTORY, physical EXAMination. This information may be important for a user, e.g., the different evidential standing of a finding reported in the history versus one established by physical examination.

Subtypes of the TREATMENT FACT type correspond to linguistic differences between statements describing general medical management (GEN), surgery (SURG), medication (MEDS), and "complementary" treatments (COMP) such as *bedrest, physiotherapy,* etc. An example of a linguistic difference among TREATMENT FACT subtypes is the presence of dosage expressions in the MEDS subtype.

The LSP System

System Architecture

The LSP information-formatting program is composed of five modules that operate in sequence on each successive sentence of the document, as illustrated in Figure 5. Equipped with a sublanguage dictionary and grammar, the program determines what statement types are present in a given sentence, creates the appropriate information formats and con-

MEDICAL CLASSES	DESCRIPTION	EXAMPLES IN ENGLISH AND FRENCH
	*** PATIENT AREA ***	
H-PT	words referring to patient	*she, le patient, elle, Mme XXX*
H-PTAREA	anatomical area	*edge, left, surface, rebord, gauche*
H-PTFUNC	physiological function	*BP, TA, appetite, tonalité, digestif*
H-PTLOC	location relation	*radiating, localisé, irradiant*
H-PTMEAS	anatomical measure	*height, size, corpulence, taille*
H-PTPART	body part	*arm, liver, bras, foie*
H-PTPALP	palpated body part	*abdomen, liver, foie*
H-PTSPEC	specimen from patient	*blood, sang, urine*
H-PTVERB	verb with patient subject	*complains of, se plaint de, subi*
	*** TEST / EXAM ***	
H-TXCLIN	clinical exam, action	*auscultation*
H-TXPROC	examination procedure	*ultrason, gastroscopie*
H-TXSPEC	test of specimen	*urine analysis*
H-TXVAR	test variable	*glucose, GB, sédiment*
	*** TREATMENT AREA ***	
H-TTGEN	general medical management	*follow-up, soins, consultation*
H-TTMED	treatment by medication	*aspirine, clamoxyl*
H-TTFREQ	frequency of medication	*bid*
H-TTMODE	mode of administration	*IM, IV*
H-TTCHIR	surgical procedures	*hysterectomy, cholécystectomie*
H-TTCOMP	complementary treatments	*bedrest, repos, physiothérapie*
	*** TIME AREA ***	
H-TMBEG	beginning	*onset, dévelope, apparition*
H-TMEND	termination	*discontinue, arrêt, stopper*
H-TMPER	duration	*persistant, constant*
H-TMREP	repetition	*habituelle, intermittent*
H-TMPREP	time preposition	*during, après, avant, depuis*
H-TMLOC	location in time	*recently, actuelle, déjà, post-op*
	*** RESULT AREA ***	
H-AMT	amount or degree	*much, totale, sévère, tout à fait*
H-BEH	behavior	*works, studies, travaille*
H-DIAG	diagnosis	*diabetes mellitus*
H-INDIC	disease indicator word	*fever, swelling, pain, thrombose*
H-NORMAL	non-problematical	*within normal limits, bon état, simple*
H-ORG	organism	*staph*
H-TXRES	test/exam result word	*positif*
H-RESP	patient response	*relief*
H-CHANGE	indication of change	*augmenté, diminution*
	*** EVIDENTIAL AREA ***	
H-NEG	negation of finding	*no, not, ne pas, jamais*
H-MODAL	uncertainty of finding	*evocatrice, probable, suspicion, semble*
	*** CONNECTIVE AREA ***	
H-BECONN	classifier verb	*is (a), est (un)*
H-CONN	P/V/ADJ/N connects two I-F's	*due to, secondaire à*
H-SHOW	V connects test and result	*shows, confirme, montre*

Figure 3 Medical subclasses in the LSP system based on word co-occurrence patterns seen in patient documents. The subclasses in the connective area are shown only in part.

nective relations, and maps the words of the sentence into the resulting structures. The LSP system for medical documents now operates in English, French, and (at the level of a PhD thesis) German.[54] The type of grammar used made it relatively straightforward to move the system from English to neighboring languages. The medical portions of the system carried over with almost no change.[55–57]

Currently the base dictionaries for the English and French versions of the LSP medical language processing system, exclusive of non-LSP supplemental sources, each contain about 10,000 words. The process of adding words to the LSP English medical dictionary is partially automated, drawing upon two large dictionaries of English words and a medical morphology program that codes the latinate words according to their internal composition.[58] Syntactic analysis of input sentences in cooperation with lexical selection rules determine the boundaries of individual fact units and the choice of appropriate I-Fs for successive sentences of input documents.

Referring to Figure 5, the first module parses the sentence into its grammatical components using a grammar that embodies syntactic structures and constraints.[32] The second module filters out alternative syntactic analyses that are not semantically correct based on established patterns of medical word-class combination (medical co-occurrence patterns or "selection lists"). The third (transformation) module makes every conjunctional substatement complete (e.g., by expanding *pain in epigastrium and right lower quadrant* to *pain in epigastrium* and *pain in right lower quadrant*) and also in other ways reduces syntactic variation. Regularization treats the connective structure, turning the whole into Polish notation. Finally, the formatting module places sentence words into the appropriate slots of the I-Fs and prepares the output for mapping into the current database structure.

While the output of the information-formatting program is in the form of tree structures, each infor-

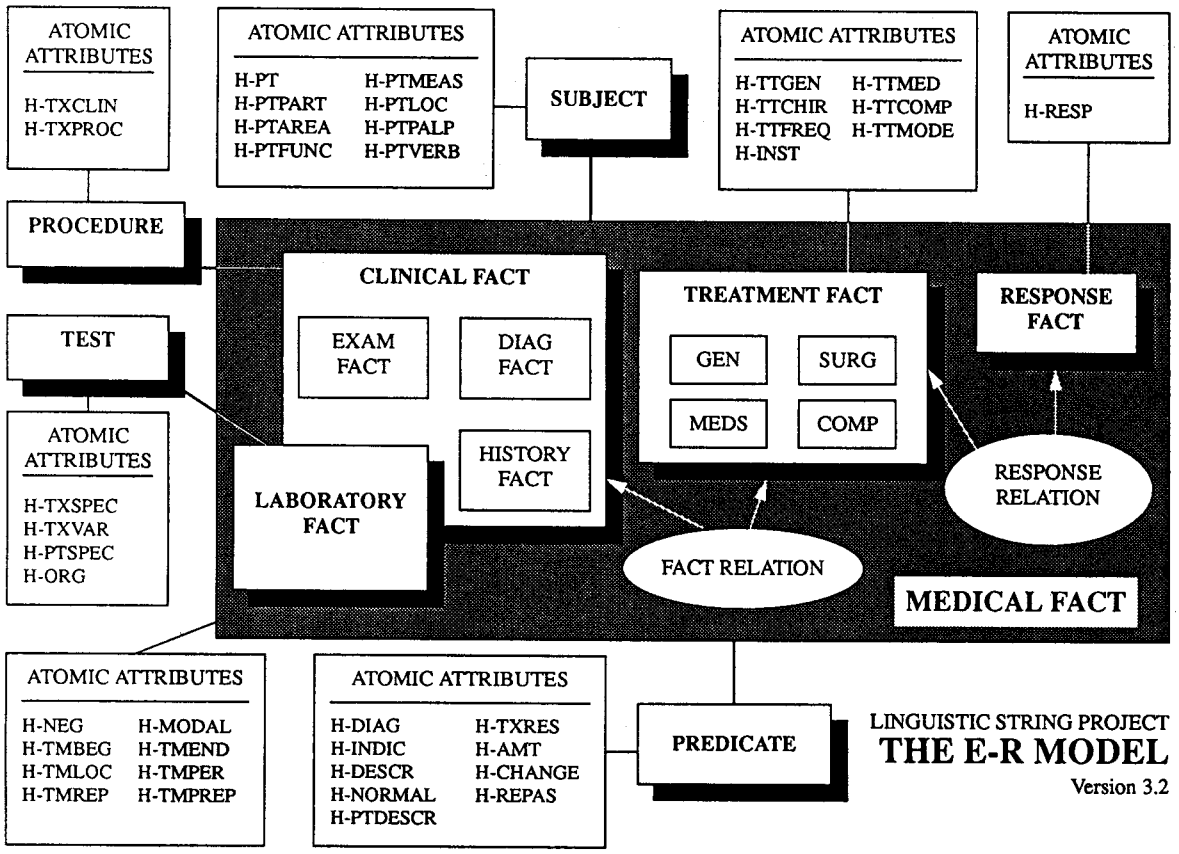

Figure 4 Schematic overview of the types of medical facts seen in patient documents and their associated lexical ("atomic") attributes. The CLINICAL FACT subtypes are distinguished by the paragraph they occur in: EXAM, DIAG, LAB, HISTORY. The TREATMENT FACT type is subdivided into general medical management (GEN), surgery (SURG), medications (MEDS), and all other therapies (COMPlementary). An instance of a TREATMENT FACT is often coupled to a RESPONSE FACT via a RESPONSE RELATION, e.g. *Much improved on penicillamine 750 mg daily.*

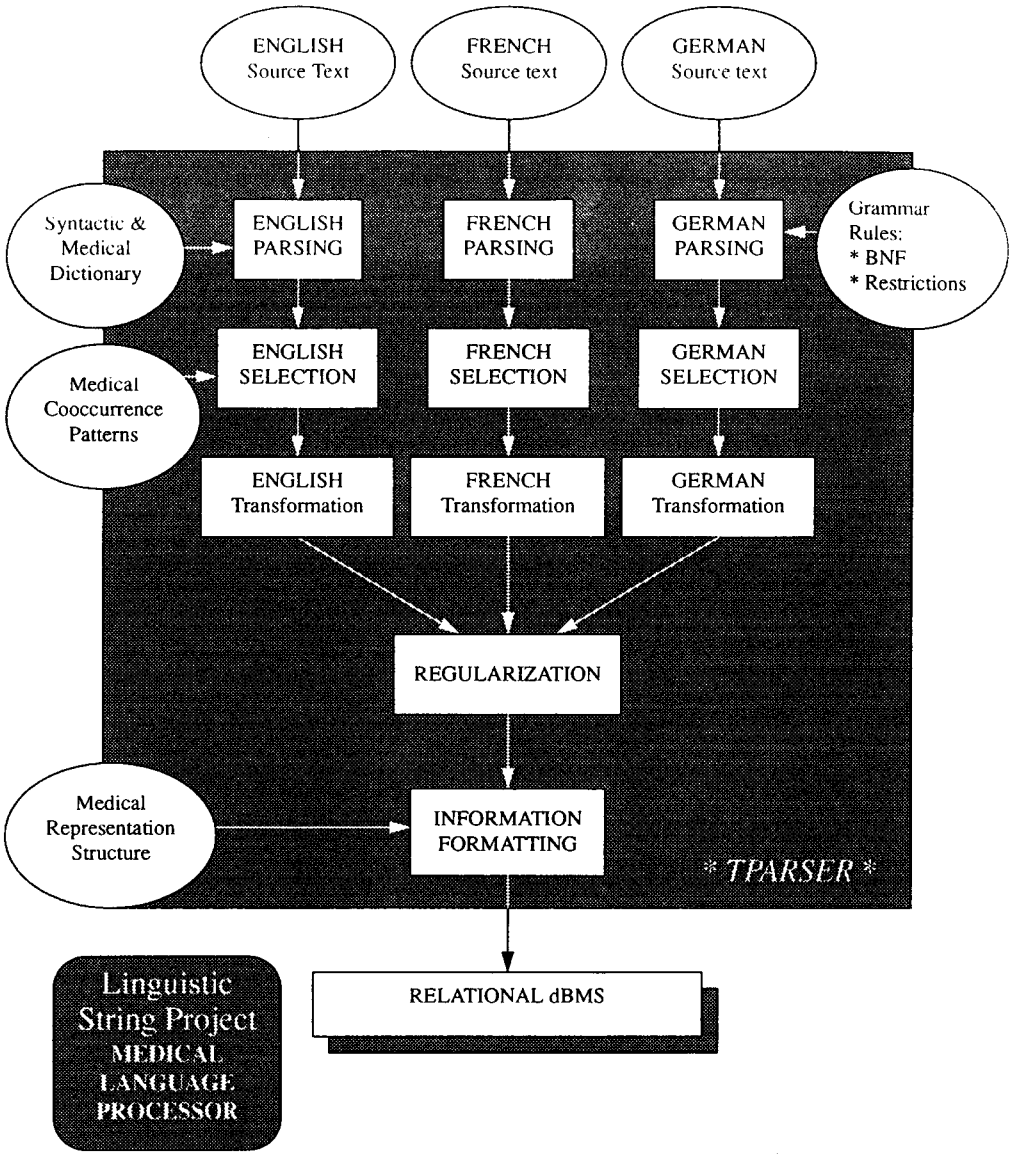

Figure 5 Schematic of the LSP language processor showing the five linguistic modules that transform the free-text input into Information-Format trees, which are then mapped to a database.

mation format tree, in the current system, is mapped into a "flattened" form to become a row of a relational database table. The growing availability of object-oriented database management systems gives hope that it will be possible to store a richer type of representation and allow for more complex retrievals without overly large user programming effort.

Quality Control of Language Processing

To ensure that medical language processing when applied to patient documents produces reliable patient data, the LSP system contains a procedure for quality control of language processing (the "NIMF" program). A database field is created to hold the results of the quality assessment of the row to be loaded. It contains:

(*empty*)	if the row passes all tests;
N	if there is a potential **N**egative problem;
I	if the row is semantically **I**ll-formed (wrong type word in field);
M	if there is a potential **M**odal problem (modal = uncertainty);
F	if there was a processing **F**ailure (the whole sentence is in the TEXTPLUS field).

Rows with a non-empty quality assessment field are not loaded into the database; the sentences with such problems are rerun using a modified parsing procedure that recovers well-formed rows from analyzable parts of sentences. All the original text material is in the database, either analyzed or in TEXTPLUS.

Quality assessment of potential database rows corresponding to an input sentence involves the lexical attributes of the words in the sentence and the fields of the database where these words should appear, given their particular lexical attributes. Thus, if a sentence contains a word with a negative attribute (H-NEG), then at least one of the rows derived from the sentence must have the NEG field filled, otherwise all rows derived from this sentence are marked **N**. Similarly, a sentence with an uncertainty word (H-MODAL) must give rise to a filled MODAL field or all its rows are marked **M**. A potential database row is marked **I** if any field contains a word with a lexical attribute that is not permitted in that field, for example, a symptom word H-INDIC in a QUANT field. The mark **F**, as noted above, applies to sentences for which no analysis was obtained.

While the NIMF program detects errors of wrong placement of words in database fields, it cannot detect an omission. Thus, in the case where a negation should be distributed over several conjoined phrases (e.g., *"no nausea, vomiting, or diarrhea"*), there is no test to assure that the NEG field is filled in all the rows created for phrases within the scope of the negation. Such a test would have to employ a language processor more powerful than the one being tested. In the LSP case, a very large emphasis has been placed on the development of the syntactic component, especially concerning conjunctions. Manual check of the processing results for all sentences containing negative or uncertainty words in a number of data sets has shown that most of the problems of these types are signaled by the NIMF program.

The recovery procedure uses segment parsing, a technique similar to the barrier-word method[59] and to procedures used in some other systems.[60] If the

LSP MLP PROCESSING REPORT
glcb DOCUMENTS

A. DOCUMENTS AND PROCESSING

Full glcb documents consist of 59 documents and 1224 sentences.
System hardware and software support from New York University, Sun Microsystem 3/50 of the Linguistic String Project, SunOs 4.1.

B. CURRENT MLP SYSTEMS
[List of grammar and dictionary files used in run]
C. SCORES OF MLP PROCESSING

Failed DICT 0 = 0.000 % of total 1224 sentences.

Of a total of 1224 processed sentences:
Failed PAR 36 = 2.941 % of 1224 processed sentences
Failed SEL 17 = 1.389 % of 1224 processed sentences
Failed XFR 6 = 0.490 % of 1224 processed sentences
Failed REG 3 = 0.245 % of 1224 processed sentences
Failed FMT 9 = 0.735 % of 1224 processed sentences
Made to dB 1153 = 94.199 % of 1224 processed sentences.

>>> MLP processing of glcb produces 2755 rows in database loadable file glcb.lodf.

D. NIMF ERROR MONITORING

>>> 266 sentences having NIMF marking are extracted from glcb, named sglcb.ocf, to be rerun with MLP segment processing for error recovery.

>>> MLP processing of glcb produces 2183/2755 rows in database loadable file glcb.lodg.

E. PROCESSING METHOD AND LOGS
[List of failures by procedure and module]
F. TRANSFORMATION FAILURES
[detailed list of failed procedures in Module 3]

LSP MLP PROCESSING REPORT
sglcb DOCUMENTS

sglcb is a list of 266 sentences marked N/M/I/F in MLP full sentence processing, reprocessed with MLP segment processing for error recovery. This is the error recovery report.

A. DOCUMENTS AND PROCESSING

sglcb document consists of 475 segments obtained from 266 sentences.
System hardware and software support from New York University, Sun Microsystem 3/50 of the Linguistic String Project, SunOs 4.1.

B. CURRENT MLP SYSTEMS
[List of grammar and dictionary files used in run]
C. SCORES OF MLP PROCESSING

Failed DICT 0 = 0.000 % of total 475 segments.

Of a total of 475 processed segments:
Failed PAR 55 = 11.579 % of 475 processed segments
Failed SEL 8 = 1.684 % of 475 processed segments
Failed XFR 6 = 1.263 % of 475 processed segments
Failed REG 1 = 0.211 % of 475 processed segments
Failed FMT 10 = 2.105 % of 475 processed segments
Made to dB 395 = 83.158 % of 475 processed segments.

>>> MLP processing of sglcb produces 672 rows in database loadable file sglcb.lodf.

D. NIMF ERROR MONITORING

>>> MLP processing of sglcb produces 298/672 rows in database loadable file sglcb.lodg.

E. PROCESSING METHOD AND LOGS
[List of failures by procedure and module]
F. TRANSFORMATION FAILURES
[detailed list of failed procedures in Module 3]

Figure 6 Excerpts from the standard report generated for a run of documents through the LSP system, shown here for the set of 59 asthma discharge letters used in the experiment described in the text.

Table 1 ■

Checklist of Important Details of Asthma Management*

1. Therapy before admission
2. Time since asthma well controlled
3. Admission
   a. Pulse
   b. Peak flow
   c. Abnormal findings
   d. Chest X-ray
   e. Blood gases
4. Treatment given including oxygen
5. Discharge
   a. Peak flow rate
   b. Result of repeat blood gases, if done
6. Arrangements at discharge
   a. Prednisolone dose and duration
   b. Long-term therapy
   c. Review arrangements

*The features of asthma management that were translated into SQL queries and applied to a database generated by the LSP system from 59 asthma discharge letters. Prepared by Dr. Christine Bucknall, Department of Respiratory Medicine, Glasgow Royal Infirmary.

processor fails to reach an analysis for a sentence (rows marked **F**) or has a potential negative or modal or semantic illformedness problem (rows marked **N, M, I**) the parser looks at a list of separators and breaks the sentence into segments at the separators. The segments are then marked sequentially from left to right and each segment is punctuated by a place holder for the separator that marks off the segment.

Segments are processed from left to right. To obtain the maximally longest analyzed segment, if a segment fails processing, it is concatenated with the next segment and its separator to form a longer segment. If the segment is the last one of the sentence, it is appended to the previous segment to make a longer segment. In case a phrase is a sentence adjunct, such as time adjuncts, then it is the parsing grammar that directs the LSP parser to add the next or previous segment for reprocessing.

Segment parsing loses connective information but on the whole does not degrade the individual fact units that are recovered. This is because the separators have been chosen with this constraint in mind. They are mainly subordinate conjunctions or prepositions (e.g., *because, because of, due to*) having the property that negation and modal markers have local scope within their boundaries. Thus, coordinate conjunctions and comma are not on the separator list. Also excluded are connectives that affect the factuality of their arguments (e.g., *without, except*).

After each run of a document set, the LSP generates a report of the processing. An excerpt from a report of the processing of the 59 discharge summaries in this study (reprocessed as one set) is shown in Figure 6. In Figure 6, section C in the left column shows the results of the LSP language processor before removal of some sentences for reprocessing. Section D in the left column shows the results of NIMF monitoring that selects sentences for reprocessing. Sections A through F in the right column show the results of reprocessing the selected sentences using segment parsing. Sentences that fail the NIMF monitoring after segment parsing are reviewed manually.

Medical Applications

Over the past four years a joint project of the LSP group and the Centre d'Informatique Hospitalier of the Hôpital Cantonal Universitaire de Genève (HCUG) has had as its goal the utilization of patient documents as a source of new clinical knowledge. HCUG has had for many years an integrated hospital information system DIOGÈNE.[61] Online text capture in DIOGÈNE is currently functioning in the departments of digestive surgery and internal medicine. The LSP system was adapted to operate on French-language discharge letters to produce a database of all the facts reported in discharge letters. When in full operation this will allow correlations among clinical findings to be investigated statistically and permit hypothesis development concerning the effects of different therapies to be explored prior to a prospective study.[50]

From early in LSP development, it seemed to us that medical quality assurance was an area that would benefit from the use of medical language processing.[62] One of the first experiments in applying the system to clinical narrative (hospital discharge summaries) was to program retrieval queries following the Performance Evaluation Procedure (PEP) of the Joint Commission for the Accreditation of Hospitals (JCAH), now the Joint Commission for the Accreditation of Health Care Organizations (JCAHO), to operate on LSP-analyzed discharge summaries.[46] For the 99 items answered by retrievals and independently by a physician reviewer, there was 91% agreement.

The investigation reported here in the area of quality assurance[63] utilizes a checklist of important details of asthma management, developed and used by Bucknall et al. of Glasgow Royal Infirmary (GRI)[64,65] (Table 1). Dr. Bucknall used this checklist in reviewing pa-

$ DOCUMENT CB07 $

$$ DOB $$
Born 00/00/33.^

$$ HISTORY $$
This 57 year woman with severe and steroid dependent asthma was admitted with an acute exacerbation which came on 48 hours before admission when she started coughing up green sputum. She has been attending the Respiratory Department for many years and is well known to Dr YYY and staff. In 00/09/90 she had local excision of an intraduct carcinoma of the left breast. Axillary nodes were clear. She is currently on Tamoxifen. Drugs on admission: Tamoxifen 20 mg daily, Nifedipine retard 10 mg bd, Duovent and Becloforte Inhalers, Prednisolone 10 mg daily and Bendrofluazide.^

$$ EXAMINATION $$
She was cushingoid and breathless with a pulse of 100, BP 150/100. She had reduced air entry and widespread wheezes. The peak flow could not be recorded. Blood gases on high flow oxygen (8 l/min) were H+ 33, PCO2 5.0, bicarb 28, PO2 24.7. The chest x-ray showed some increased shadowing at the left base.^

$$ INVESTIGATION & PROGRESS $$
She was treated with antibiotics, increased steroids and nebulised bronchodilators. She made a gradual but slow recovery. Her peak flow took 48 hours to come up to 100 and her best peak flow was 220 at discharge. While on the ward she complained of persistent lumbar back pain. A lumbar spine x-ray and isotope bone scan failed to reveal any abnormality and her local discomfort was probably due to persistent coughing.^

$$ DISCHARGE $$
Discharged 11/04/91. The drugs on discharge: Tamoxifen 20 mg daily, Nifedipine Retard 10 bd, Prednisolone 20 mg daily, Bendrofluazide 5 mg daily, Theophylline 250 mg at night, Co-Codamol. She will be seen again in the Respiratory Clinic in two weeks.^

Figure 7 A sample patient document from the Glasgow Royal Infirmary after preprocessing for input to the LSP system.

tient records, and used a questionnaire with home visit to assess outcome. Use of the LSP system was initiated to see if computer analyses of asthma discharge letters (or summaries) could provide data needed to convert the periodic evaluation process to a continuous medical audit.

Document Processing

A typical discharge letter from GRI with some preprocessing steps already completed is shown in Figure 7. Before the documents were entered into the computer, a series of operations was applied to the text to indicate uniformity in:

1. Paragraph headings (taking the most commonly used paragraph headings in the document set and applying them to all documents). They are labeled for processing by "$$".
2. Date expressions (here following the United Kingdom usage of day/month/year).
3. Names: XXX for patient names, YYY for doctors' names, and ZZZ for hospital, clinic, or other location names.
4. Paragraph endings (^), use of periods as sentence terminator except in numerical expressions and selected abbreviations, and use of space to delineate words or symbols.
5. Document identification number and name of document.

Documents were processed using the LSP system, with the information-formatted data mapped into a relational database. Retrieval programs written in Structured Query Language (SQL) were applied to the database to retrieve the information required for the checklist. An initial set of 28 asthma discharge letters was used to train the system (i.e., update the LSP English medical dictionary, modify the grammar as needed, and develop the SQL queries). A second set of documents was then used to test system per-

REPRESENTATIVE RETRIEVALS FOR CHECKLIST ITEMS 2 AND 3c, WITH TEXT

2. TIME SINCE ASTHMA WELL CONTROLLED

SID	ROW	SIGN-SYMPTOMS SUGGESTING ASTHMA	QUANT	TIME WORDS
002B.1.01	1	OF EXERTIONAL WHEEZY DYSPNOEA	STEADILY INCREASING	A 4 DAY HISTORY
002B.1.02	2	OF DETERIORATING WHEEZY BREATHLESSNESS	ACUTELY	A 4 TO 5 HOUR HISTORY
046B.1.01	2	BREATHLESSNESS	INCREASING	OVER THE FEW DAYS PREVIOUSLY
047B.1.02A	1	BREATHLESS	INCREASINGLY	IN THE FEW HOURS PRIOR TO ADMISSION
056B.1.01	2	OF DYSPNOEA		A 3 WEEK HISTORY

TEXT:

002B.1.01 This 29 year old girl who is known to be asthmatic for 15 years presented with a 4 day history of steadily increasing exertional wheezy dyspnoea with a cough productive of green sputum.

002B.1.02 She was given a reducing dose of prednisolone prescribed by her own doctor; however she presented with a 4 to 5 hour history of acutely deteriorating wheezy breathlessness.

046B.1.01 Your patient was admitted as an emergency complaining of increasing breathlessness over the few days previously.

047B.1.02 She had become increasingly breathless in the few hours prior to admission though she had no problems with cough or sputum production.

056B.1.01 This 40 year old man presented with a 3 week history of dyspnoea, cough productive of white sputum and left sided chest pain.

3c. ASTHMA - ABNORMAL PHYSICAL FINDINGS AT ADMISSION

SID	ROW	SIGN-SYMPTOM OR DIAG	QUANT	BODY PART	BODY FUNCTION
002C.1.01	1	DROWSY			
002C.1.01	2	SMELLING OF ALCOHOL	STRONGLY		
002C.1.05	1	VESICULAR		OF THE CHEST OF HER CHEST	BREATH SOUNDS IN BOTH SIDES
002C.1.05	2	RHONCHI	WIDESPREAD		EXPIRATORY
046C.1.01	3	WHEEZY	MARKEDLY		
046C.1.07	1	DIFFUSE RHONCHI		CHEST	EXPIRATORY
046C.1.07	2	CREPS		BI-BASAL	
046C.1.08	1	OBESE		ABDOMINAL	
047C.1.02	2	WHEEZING	QUITE MARKED	OF THE CHEST	RESPIRATORY
056C.1.01	1	TACHYPNOEIC			
056C.1.03B	1	HYPERINFLATED		CHEST	
056C.1.04	1	A RUB IN THE MIDZONE		PLEURAL LEFT	
056C.1.05	1	WITH SCARS		ABDOMEN	

TEXT:

002C.1.01 On examination she was drowsy, smelling strongly of alcohol.

002C.1.05 On examination of the chest she had vesicular breath sounds present in both sides of her chest with widespread expiratory rhonchi but no focal signs of infection.

046C.1.01 Examination on admission revealed a 56 year old woman who was apyrexial and markedly wheezy.

046C.1.07 Chest: Diffuse expiratory rhonchi with bi-basal creps.

046C.1.08 Abdominal examination: obese.

047C.1.02 Examination of the cardiovascular system was apparently unremarkable though auscultation of the chest revealed quite marked respiratory wheezing.

056C.1.01 He was tachypnoeic.

056C.1.03 No oedema, chest hyperinflated, there was widespread rhonchi throughout.

056C.1.04 There was a possible pleural rub in the left midzone.

056C.1.05 Abdomen was soft with previous surgical scars.

Figure 8 Representative retrievals for asthma checklist items 2 and 3c (time since asthma well controlled, abnormal findings at admission) for four patients, along with the contributing text.

Table 2 ■

Checklist of Important Details of Asthma Management*

	Number of Documents (Total 28. Training Set)				Number of Documents (Total 31. Test Set)			
	Information Not Present	Information Present; Retrieved Correctly	Information Present; Retrieved with Some Errors†	Information Present; Missed in Part or Whole‡	Information Not Present	Information Present; Retrieved Correctly	Information Present; Retrieved with Some Errors†	Information Present; Missed in Part or Whole‡
1. Therapy before admission	7	17	(1,0)	(2,1)	8	14	(4,1)	(3,1)
2. Time since asthma well controlled	10	15	(0,0)	(0,3)	10	18	(0,0)	(0,3)
3. Admission								
a. Pulse	7	21	(0,0)	(0,0)	9	21	(0,1)	(0,0)
b. Peak flow rate	15	8	(2,2)	(0,1)	12	9	(6,2)	(1,1)
c. Abnormal findings	1	21	(6,0)	(0,0)	2	20	(5,0)	(3,1)
d. Chest X-rays	10	16	(1,0)	(1,0)	9	10	(2,0)	(3,7)
e. Blood gases	8	12	(3,0)	(4,1)	9	17	(0,0)	(4,1)
4. Treatment given including oxygen	0	20	(4,2)	(0,2)	0	18	(12,0)	(1,0)
5. Discharge								
a. Peak flow	22	5	(0,0)	(1,0)	20	10	(0,0)	(0,1)
b. Repeat blood gases, if done	22	6	(0,0)	(0,0)	30	0	(0,0)	(0,1)
6. Arrangements at discharge								
a. Prednisolone	8	15	(1,0)	(1,3)	7	17	(2,0)	(3,2)
b. Long-term therapy	4	20	(0,0)	(2,2)	3	23	(3,0)	(0,2)
c. Review arrangement	3	18	(0,0)	(5,2)	3	15	(9,0)	(0,4)

*Summary of retrieval results for queries corresponding to the checklist of important details of asthma management, applied to a database generated by LSP processing of 59 discharge letters.
†In this column (n_1, n_2): n_1 = minor error, n_2 = major error.
‡In this column (n_1, n_2): n_1 = minor miss, n_2 = major miss.

formance, after updating the dictionary for text words not yet in the dictionary.

Retrieval Results

This section shows the results of retrievals for the checklist items deemed important in asthma management (Table 1) as translated into SQL queries and applied to the relational database created by the LSP system from the 59 GRI discharge letters. Tables of retrieved information obtained for each query were compared with the text of the original documents by physician reviewers (Drs. Bucknall and Lyman).

Figure 8 shows, for four patients (Patients 2, 46, 47, and 56), the sections of the tables generated by SQL queries for checklist items 2 (time since asthma well controlled) and 3c (abnormal physical findings at admission) along with the original contributing sentences, shown for clarity. Item 2 was interpreted to mean that specific time information was present in the document with regard to at least one of the presenting symptoms (e.g., 056B.1.01). Item 3c was evaluated with respect to the report of the admission physical examination. Since several sentences often were involved, all findings from all sentences were considered. An error was counted as minor if a misplaced word did not destroy the main meaning; a major error could be retrieving as an admission finding one that was not found at admission. A "miss" was considered minor if two-thirds of the findings were retrieved (e.g., in Fig. 8, retrieval 3c for patient 56 returned four abnormal findings out of five; in 056C.1.03 it missed *rhonchi*), and major otherwise.

Table 2 provides the results from the two sets of documents, using the document as the unit of measurement. The columns, reading across, are:

1 and 5: information not present in the document;
2 and 6: information present in the document and retrieved correctly;
3 and 7: information present and retrieved with minor or major error;

4 and 8: information present and retrieved with minor or major portions of the information missing.

Thus, for checklist item 1 (therapy before admission), seven of 28 discharge letters in the training set did not contain this information (column 1), nor did eight of 31 in the test set (column 5). The LSP system in conjunction with SQL queries correctly retrieved all such information for this query from 17 of 28 discharge letters in the training set (column 2) and 14 of 31 in the test set (column 6). Columns 3 and 7 show some errors in the reports of information for each set: one minor and no major for query 1 applied to the training set and four minor and one major for query 1 applied to the test set of discharge letters. Some or all of the desired information for this query was missed in three discharge letters of the training set (column 4), of which two were minor misses and one was major; information was missed in four documents of the test set (column 8), of which three were minor misses and one was major. With the simple retrieval queries in this experiment, the system did not retrieve data items that were totally unrelated to the query (false positives).

Only columns 2, 3, 4 for the training set and 6, 7, 8 for the test set are involved in evaluation of the performance of the computer processing (NLP + SQL queries). We defined information precision (I-P) for each query addressed to the set of documents under study:

$$\text{I-P} = \frac{\text{the number of documents for which the desired information was retrieved}}{\text{the total number of documents for which any information was retrieved}}$$

Similarly, we defined information recall (I-R) for each query addressed to the document set:

$$\text{I-R} = \frac{\text{the number of documents for which the desired information was retrieved}}{\text{the total number of documents that contained such information}}$$

Table 3 shows information precision (I-P) to be 91.1% for the 13 queries applied to the database of 28 computer-analyzed discharge letters (training set) and 82.1% for the test set of 31 documents. When I-P was calculated using the counts of major errors only, the scores were 95.7% for the training set and 98.6% for the test set. Information recall (I-R) for the 13 queries applied to the database of 28 analyzed discharge letters (training set) was 86.9%; it was 82.5% for the test set of documents. When I-R was calculated using counts of major omissions only, the corresponding scores were 93.9% for the training set and 92.5% for the test set.

Discussion

The performance of the language processor is discussed separately from the results of the retrievals. The two components are almost in the elephant–flea relation: the language component is the result of a long-term development; the SQL retrievals were written (without specialized database expertise) primarily to test the language-processing system.

Language-processing Results

Figure 6 provides some performance data for the LSP language processor, not including the retrieval component, from a rerun of the 59 asthma documents (training set plus test set). The success rate of 94.2% sentences ("Made to dB" in the report shown in Fig. 6) is rather higher than one would usually expect. The processing of several sets of French Lettres de Sorties (ca. 7,000 sentences) was performed with a global success rate of about 85%.[66] Performance figures are for complete documents, including paragraphs of discussion and evaluation that rarely contribute to the retrievals in the applications considered to date and are linguistically considerably more complex. Diagnosis and physical examination paragraphs score much higher than the average for all paragraphs. History paragraphs score relatively high except for some initial sentences in which the physician appears to be trying to say it all in just one sentence.

The question is often raised as to the timing of the processes described here in anticipation of possible practical applications. It would be important to know whether a reasonable machine in a reasonable time

could complete the tasks. It is possible to give a number that is probably correct to an order of magnitude for extrapolation purposes. These texts were run on a SUN Sparcstation 1 with 16 mb of memory. This machine has a local swapping disk but the data reside on an NFS mounted server on a very busy-to-moderately busy network. The source program (FORTRAN) has been carefully optimized, but it was not compiled with the optimization option turned on—this option has been known to produce improvements of 10–20%, but can create problems. More importantly, for a considerable part of the running time there were other users on the machine. The parser, the dictionary, and the five processor components were loaded and unloaded many times (66 times for the 59 asthma documents).

Another factor in the timing results is the fact that the grammar procedures are heavily loaded with diagnostics (the source material for the generated report). As would be expected, the speed of processing depends on sentence complexity. The time may be immediate (much less than a second) or in some cases run to minutes. With these caveats we give some timings: 14 seconds/sentence for the 59 asthma documents (1,224 sentences); 19 seconds/sentence for a set of French Lettres de Sorties (1,442 sentences).

A knowledge of the sources of error is important. The most difficult part of the processing is the parsing phase. Of the roughly 10–15% overall failure rate for sentences not reaching the database on the first pass, most failures are in the parsing module. This is partly by design, in the sense that most of the effort to prevent wrong analyses is concentrated in constraints on the parsing process. The constraints include grammatical rules (adapted for the quaint style of clinical reporting) and many semantic checks on the kinds of words that can fill particular syntactic roles in relation to other words in the same structure. The semantic checks depend for their effectiveness on the correctness and consistency of the classification of words into medical classes, i.e., the quality of the dictionary in relation to the requirements of the rules in the parsing component (and secondarily the requirement of the selection component).

A failure in any particular procedure in the four modules that follow the parsing is recorded for the detailed report that is generated for every LSP pro-

Table 3 ■

Some Performance Measures of the LSP System and SQL Retrievals in Extracting Quality Assurance Information from 59 Discharge Letters

	Training Set (28 Documents)		Test Set (31 Documents)	
Checklist Item	Information Precision (%)	Information Recall (%)	Information Precision (%)	Information Recall (%)
1. Therapy before admission	94.4	85.0	73.6	77.8
2. Time since asthma well controlled	100.0	83.3	100.0	85.7
3. Admission				
a. Pulse	100.0	100.0	95.4	100.0
b. Peak flow rate	66.6	88.8	52.9	81.8
c. Abnormal findings	77.7	100.0	80.0	83.3
d. Chest X-rays	94.1	94.1	83.3	50.0
e. Blood gases	80.0	70.6	100.0	77.3
4. Treatment given including oxygen	77.7	90.9	60.0	94.7
5. Discharge				
a. Peak flow	100.0	83.3	100.0	90.9
b. Repeat blood gases, if done	100.0	100.0	*	†
6. Arrangements at discharge		78.9	89.5	77.3
a. Prednisolone	93.7	83.3	88.5	92.0
b. Long-term therapy	100.0	72.0	62.5	78.9
c. Review arrangement	100.0			
AVERAGE	91.1	86.9	82.1	82.5
Average major errors/misses only	95.7	93.9	98.6	92.5

*Number of errors/number retrieved = 0/0.
†Number should be retrieved = 1, too small for calculation.

cessing job. Various reporting options that trace the operation of the parsing process are available on demand.

Evaluation and Utility of Retrievals

New evaluation measures will have to be developed for situations involving complex information retrieval, such as the extraction of information from written text. The variables include whether a document containing the information was located (appeared in retrieval results), whether the information was represented correctly in the retrieval results, and whether the retrieved information was complete. Major as opposed to minor departures from total correctness will also have to be dealt with in relation to the goals of the application. As a start, for this study we defined the measures information precision (I-P) and information recall (I-R), as above, adapted from the established definitions of precision and recall used in bibliographic retrieval. Investigators comparing the overlap of the UMLS Metathesaurus with the diagnostic information in the Iliad expert system also had to define new metrics. They distinguished four types of matches and three types of non-matches.[67]

In Table 2, columns 1 and 5 provide data describing deficits in the quality of documentation of patient care with respect to stated criteria. If one considers the items not mentioned to be important in asthma management, one can see immediately where emphasis on teaching could be placed, or feedback to the reporting physicians introduced. It should be noted that with regard to checklist Item 5b, "repeat blood gases, if done," the SQL query counted all cases of no second mention of blood gases as the entry for column 1, illustrating how carefully queries must be formulated and how cautiously computer results should be treated as a reflection of physician deficit in documentation. There may well be reasons why suggested procedures are not performed, as the text in several of the GRI documents testified.

We view the NLP techniques when used for quality assurance to be in the nature of a tool for screening clinical documents with respect to quality-of-care criteria, not as an automated evaluator of physician performance. In the real world of audit, knowing which discharge summaries had significant deficits in documentation would allow the human reviewer to be selective with regard to which records need manual review. Other uses of NLP for quality assurance (using another NLP system) are also being investigated.[68]

In terms of the potential use of language processing as an aid in the audit task, the language-processing tool demonstrated here has several significant features:

1. All documents are treated consistently with regard to a given criterion.
2. Only significant information is retrieved.
3. Major errors in retrieval results are minimal (average 1.4% for the test set in this experiment).
4. Major omissions in the retrieval results are relatively small in number (average 7.5% for the test set in this experiment).
5. The semantic structuring and relative completeness of retrieved data suggest their potential use as input to further quality-assurance procedures.

NLP, Vocabularies, and Coding

The "end point" of the process described here is a set of rows of a relational database whose fields contain text. As is often the case in development efforts, many choices were influenced by available resources. We have indicated that the database management system used so far (RDBMS) has severe limitations and probably should be replaced with improved products (OODBMS) as they become available. Also, there is no requirement that this type of endpoint be the "end." The history of computer-based systems shows layering as a common feature. There is no barrier to mapping the sort of database described here to one that has a more standardized terminology or even codes.

Given a standard vocabulary, one may hope to use a medical language processor (the one described here or others that are being developed[60,69,70]) for automatic encoding of narrative patient documents. This is not an easy task, because automatic encoding implies that there is a commonality to the expressions in the text to be coded and the word strings of the code, which may be difficult to establish. In part this difficulty arises because the text is saying more about the entities in question than that they were mentioned, and in part because the mentions themselves may appear in varied forms. The existence of a faceted vocabulary, as is the case with SNOMED, or semantic typing as in the UMLS Metathesaurus development, provides lexical classifications that can facilitate the matching, but it is important that the categories be applied consistently within the vocabularies and again to the texts via their parsing.

As the experiment in asthma quality assurance reported here illustrates, NLP techniques can be used within a specific application that requires data in part

contained in narrative patient documents. NLP may also be useful in conjunction with established vocabularies to provide a linguistic underpinning to the development of standard representations of clinical data. It is possible that the use of linguistic analysis can help in making vocabularies more internally consistent and provide the basis for user tools to navigate among related expressions. There is work in progress on SNOMED III related to such goals, using conceptual graphs as the formalism[71,72] and using frames.[73]

Conclusion

This paper describes a system for processing the narrative portions of patient records and mapping the information elements into a database representation. A demonstration of the use of the system for medical quality assurance is presented.

The system has been under development for some years. The methodology is primarily linguistic analysis. The grammar component is relatively elaborate; the semantic component is based on the types of statements found to be characteristic in narrative clinical reports. A strict control on the quality of the language processing is in effect; overall about 85% of the sentences of input documents reach the database after passing through the quality-control program.

The system is application-independent in the sense that the processor and the major components (grammars, lexicon) are geared to the general features of clinical reporting found in most documents. Nevertheless, a period of adaptation for any particular application will be necessary. The amount of effort required to develop an application that uses the system depends very heavily on the complexity of the information that is desired from the documents and how tolerant (or intolerant) of error these demands are.

It would be misreading our results to predict that the LSP system (or any natural-language processing system) will quickly (or ever) become a simple off-the-shelf all-purpose medical language processing product. Nevertheless, a judicious use in appropriate applications may prove to be a valuable adjunct to existing clinical data collection. Hopefully, a step-by-step incorporation of language processing into the repertoire of data-processing tools will improve the access and utilization of patient data for medical quality assurance, the exploration of clinical hypotheses, and ultimately physician support in the health care process.

References ■

1. Dick RS, Steen EB, eds. The Computer-Based Patient Record. An Essential Technology for Health Care. Washington, DC: National Academy Press, 1991.
2. Ball MJ, Collen MF, eds. Computers in Health Care Series: Aspects of the Computer-based Patient Record. New York: Springer-Verlag, 1992.
3. Pratt AW. Medicine, Computers, and Linguistics. Adv Biomed Eng. 1973;3:97–140.
4. Systematized Nomenclature of Pathology. Skokie, IL: College of American Pathologists, 1976.
5. Coté RA, ed. Systematized Nomenclature of Medicine. Skokie, IL: College of American Pathologists, 1982.
6. Coté RA, Rothwell DJ, Beckette R, Palotay J, eds. SNOMED International. Northfield, IL: College of American Pathologists, 1993.
7. Dunham GS, Pacak MG, Pratt AW. Automatic indexing of pathology data. J Am Soc Inform Sci. 1978;29:81–90.
8. Wingert F, Rothwell D, Coté RA. Automated indexing into SNOMED and ICD. In: Scherrer J-R, Coté RA, Mandil S, eds. Computerized Natural Medical Language Processing for Knowledge Representation. Amsterdam, The Netherlands: Elsevier Science, 1989:201–39.
9. Coté RA, Protti DJ, Scherrer J-R, eds. Role of Informatics in Health Data Coding and Classification Systems. Amsterdam, The Netherlands: Elsevier Science, 1985.
10. Humphreys BL, Lindberg DAB. Building the Unified Medical Language System. Proc Annu Symp Computer Applications in Medical Care. 1989;13:475–80.
11. Tuttle MS, Sheretz MS, Erlbaum NE, Olson N, Nelson S. Implementing META-1*: the first version of the UMLS METATHESAURUS. Proc Annu Symp Computer Applications in Medical Care. 1989;13:483–7.
12. Cimino JJ. Representation of clinical laboratory terminology in the Unified Medical Language System. Proc Annu Symp Computer Applications in Medical Care. 1991;15:199–203.
13. Vries JK, Marshalek B, d'Arbarno JC, Yount RJ, Dunner LL. An automated indexing system utilizing semantic net expansion. Comput Biomed Res. 1992;25:153–67.
14. Medical Archival System (MARS): large scale medical data archiving at the University of Pittsburgh. Pittsburgh, PA: University of Pittsburgh, Office of Biomedical Informatics, 1990.
15. Chute CG, Yang Y, Evans DA. Latent semantic indexing of medical diagnoses using UMLS semantic structures. Proc Annu Symp Computer Applications in Medical Care. 1991; 15:185–9.
16. Gabrielli ER. Computer assisted assessment of patient care in the hospital. J Med Syst. 1988;12:135.
17. Gabrielli ER. Computerizing text from office records. MD Comput. 1987;4:44.
18. Clancy WJ, Shortliffe EH, eds. Readings in Medical Artificial Intelligence. The First Decade. Reading, MA: Addison-Wesley, 1984.
19. Blum BI, Duncan K, eds. A History of Medical Informatics. New York: ACM Press, 1990.
20. Miller RA, Pople HE Jr, Myers JD. Internist-I: an experimental computer-based diagnostic consultant for general internal medicine. N Engl J Med. 1982;307:468–76.
21. Buchanan BG, Shortliffe EH, eds. Rule-based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project. Reading, MA: Addison-Wesley, 1984.
22. Reggia JA, Tuhrim S, eds. Computer Assisted Medical Decision-Making. Vols. 1 and 2. New York: Springer-Verlag, 1985.
23. Shortliffe EH. Clinical decision-support systems. In: Shortliffe EH, Perreault LE, eds. Medical Informatics: Computer Appli-

cations in Medical Care. Reading, MA: Addison-Wesley, 1990:466–502.
24. Miller RE, Masarie FE, Myers JD. Quick Medical Reference (QMR) for diagnostic assistance. MD Comput. 1986;3:34.
25. Barnett GO, Cimino JJ, Hupp JA, Hoffer EP. DXplain—an evolving diagnostic decision support system. JAMA. 1987;258:67–74.
26. Pryor TA, Gardner RM, Clayton PD, Warner HR. The HELP System. J Med Syst. 1983;7:2.
27. Musen MA. Dimensions of knowledge sharing and reuse. Comput Biomed Res. 1992;25:435–67.
28. Masarie FE, 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.
29. Sager N. Syntactic analysis of natural language. In: Advances in Computers. Vol. 8. New York: Academic Press, 1967: 153–88.
30. Grishman R, Sager N, Raze C, Bookchin B. The linguistic string parser. In: AFIPS Conference Proceedings. Montvale, NJ: AFIPS Press, 1973;42:427–34.
31. Sager N, Grishman R. The restriction language for computer grammars of natural language. Communications of the ACM. 1975;18:390–400.
32. Sager N. Natural Language Information Processing: A Computer Grammar of English and Its Applications. Reading, MA: Addison-Wesley, 1981.
33. Harris ZS. Mathematical Structures of Language. New York: Wiley Interscience, 1968:Section 5.9.
34. Harris ZS. Language and Information. Bampton Lectures in America 28. New York: Columbia University Press, 1988.
35. Harris Z. A Theory of Language and Information: A Mathematical Approach. New York: Oxford University Press, 1991.
36. Kittredge R, Lehrberger J. Sublanguage: Studies of Language in Restricted Semantic Domains. Berlin, Germany: Walter de Gruyter, 1982.
37. Grishman R, Kittredge R, eds. Analyzing Language in Restricted Domains: Sublanguage Description and Processing. Hillsdale, NJ: Lawrence Erlbaum, 1986.
38. Marsh E, Froscher J, Grishman R, Hamburger H, Bachenko J. Automatic Processing of Navy Message Narrative. Washington DC: Naval Research Laboratory, 1985:NRL Report 8893.
39. Hirschman L, Grishman R, Sager N. Grammatically-based automatic word class formation. Inform Proc Manage. 1975:11: 39–57.
40. Sager N. Natural language information formatting: the automatic conversion of texts to a structured data base. In: Advances in Computers. New York: Academic Press, 1978;17:89–162.
41. Harris Z, Gottfried M, Ryckman T, et al. The Form of Information in Science: A Test-Case in Immunology. Dordrecht, The Netherlands: Kluwer Academic, 1989.
42. Sager N. Syntactic formatting of science information. In: AFIPS Conference Proceedings. Montvale, NJ: AFIPS Press, 1972; 41:791–800. Reprinted in: Kittredge R, Lehrberger J, eds. Sublanguage: Studies of Language in Restricted Semantic Domains. Berlin, Germany: Walter de Gruyter, 1982:9–26.
43. Sager N. Sublanguage grammars in science information processing. J Am Soc Inform Sci. 1975;10–6.
44. Sager N, Friedman C, Lyman MS. Medical Language Processing: Computer Management of Narrative Data. Reading, MA: Addison-Wesley, 1987.
45. Grishman R, Hirschman L. Question answering from natural language medical data bases. Artif Intell. 1978;11:25–43.
46. Hirschman L, Story G, Marsh E, Lyman M, Sager N. An experiment in automated health care evaluation from narrative medical records. Comput Biomed Res. 1981;14:447–63.
47. Lyman M, Chi EC, Sager N, Tick LJ, Story GA. Automated case review of acute bacterial meningitis of childhood. In: MEDINFO83: Proceedings of the Fifth International Conference on Medical Informatics. Amsterdam, The Netherlands: Elsevier Science, 1983.
48. Sager N, Wong R. Developing a database from free-text clinical data. J Clin Comput. 1983;XI,5 & 6:184–94.
49. Chi EC, Lyman MS, Sager N, Friedman C, Macleod C. A database of computer-structured narrative: methods of computing complex relations. Proc Annu Symp Computer Applications in Medical Care. 1985;9:221–6.
50. Borst F, Lyman MS, Nhàn NT, Tick LJ, Sager N, Scherrer J-R. TEXTINFO: a tool for automatic determination of patient clinical profiles using text analysis. Proc Annu Symp Computer Applications in Medical Care. 1991;15:63–7.
51. Hirschman L. Retrieving time information from natural language texts. In: Oddy RN, Robertson SE, Van Rijsbergen CJ, Williams P, eds. Information Retrieval Research. London, England: Butterworths, 1981:154–71.
52. Hirschman L, Story G. Representing implicit and explicit time relations in narrative. Proc Seventh Int Joint Conf Artif Intell. 1981;7:289–95.
53. Batini C, Ceri S, Navathe SB. Conceptual database design: an entity-relationship approach. Redwood City, CA: Benjamin/ Cummings, 1992.
54. Oliver NC. A Sublanguage Based Medical Language Processing System for German. PhD Thesis. New York: New York University, Department of Computer Science, 1992.
55. Sager N, Lyman M, Tick LJ, et al. Adapting a medical language processor from English to French. In: MEDINFO89: proceedings of the Sixth International Conference on Medical Informatics. Amsterdam, The Netherlands: Elsevier Science, 1989:548–53.
56. Nhàn NT, Sager N, Lyman M, Tick LJ, Borst F, Su Y. A medical language processor for two Indo-European languages. Proc Annu Symp Computer Applications in Medical Care. 1989;13:554–8.
57. Borst F, Sager N, Nhàn NT, et al. Analyse automatique de comptes rendus d'hospitalisation. In: Degoulet P, Stephan J-C, Venot A, Yvon P-J, redacteurs. Informatique et Santé, Informatique et Gestion des Unités de Soins, Comptes Rendus du Colloque AIM-IF, Paris, 1989. Paris, France: Springer-Verlag, 1989:246–56.
58. Wolff, S. The use of morphosemantic regularities in the medical vocabulary for automatic lexical coding. Meth Inform Med. 1984;23:195–203.
59. Moore GW, Miller RE, Hutchins GM. Indexing by MeSH titles of natural language pathology phrases identified on first encounter using the barrier word method. In: Scherrer J-R, Coté RA, Mandil S, eds. Computerized Natural Medical Language Processing for Knowledge Representation. Amsterdam, The Netherlands: Elsevier Science, 1989:29–39.
60. Lin R, Lenert L, Middleton B, Shiffman S. A free-text processing system to capture physical findings: Canonical Phrase Identification System (CAPIS). Proc Annu Symp Computer Applications in Medical Care. 1991;15:843–7.
61. DIOGENE Staff. The DIOGENE Hospital Information System. Division Informatique, Hôpital Cantonal Universitaire de Genève, 1211 Genève 4, Switzerland, 1986.
62. Sager N, Lyman M. Computerized language processing: implications for health care evaluation. Med Rec News. 1978;49: 20–30.
63. Lyman M, Sager N, Nhàn NT, Tick LJ, Borst F, Scherrer JR. The application of natural-language processing to healthcare quality assessment. Med Decis Making. 1991;11:suppl 4:S65–

S68.
64. Bucknall CE, Robertson C, Moran F, Stevenson RD. Differences in hospital asthma management. Lancet. 1988;1:748–50.
65. Bucknall CE, Robertson C, Moran F, Stevenson RD. Management of asthma in hospital: a prospective audit. Br Med J. 1988;296:1637–9.
66. Scherrer, J-R, et al. Automatic encoding of clinical narratives. In: Rapport Scientifique Intermediare, Centre d'Informatique Hospitalier, Hôpital Cantonal Universitaire de Genève, February 14, 1992.
67. Bouhaddou O, Warner H, Huff S, et al. Evaluating how the UML Metal.1 covers disease information contained in a diagnostic expert system (Iliad). Presentation at AMIA Spring Congress, St. Louis, MO, 1993.
68. Zingmond D, Lenert LA. Monitoring free-text data using medical language processing. Comput Biomed Res. 1993;26:467–81.
69. McCray AT. Extending a natural language parser with UMLS knowledge. Proc Annu Symp Computer Applications in Medical Care. 1991;15:194–8.
70. Baud RH, Rassinoux A-M, Scherrer J-R. Natural language processing and semantical representation of medical texts. Meth Inform Med. 1992;31:117–25.
71. Sowa JF. Conceptual Structures: Information Processing in Mind and Machine. Reading, MA: Addison-Wesley, 1984.
72. Campbell KE, Musen MA. Representation of clinical data using SNOMED III and conceptual graphs. Proc Annu Symp Computer Applications in Medical Care. 1992;16:354–8.
73. Evans DA, Rothwell DJ, Monarch IA, Lefferts RG, Coté RA. Toward representations for medical concepts. Med Decis Making. 1991;11:suppl 4:s102–s108.