ANNALS OF THE NEW YORK ACADEMY OF SCIENCES

Volume 583

EDITORIAL STAFF

Executive Editor
BILL BOLAND

Managing Editor
JUSTINE CULLINAN

Associate Editor
JOYCE HITCHCOCK

The New York Academy of Sciences
2 East 63rd Street
New York, New York 10021

ANNALS OF THE NEW YORK ACADEMY OF SCIENCES
Volume 583

THE USES OF LINGUISTICS

Edited by Edward H. Bendix

*The New York Academy of Sciences
New York, New York
1990*

Computer Analysis of Sublanguage Information Structures

NAOMI SAGER
Courant Institute of Mathematical Sciences
New York University
New York, New York 10003

INTRODUCTION

This paper describes results in developing computer programs for processing the information in texts. It describes, in particular, a natural language processor developed by the Linguistic String Project (LSP) of New York University. The processor accepts narrative input in particular subject areas and produces from it structured data suitable for selective retrieval and further information processing. The computer processing transforms, but does not change, the content of the texts.

One may ask what purpose is served by a text analysis capability of computers. The ability to "digest" the information in accumulating files could help to reduce the situation termed information overload. For example, in the medical domain, a single patient chart in a chronic disease situation can be several volumes. The busy clinic physician cannot read the whole record. As a result, the physician often can only treat the immediate problem. By having the record computerized, analyzed and organized, significant features of the disease process in the patient could be reviewed and the basis for therapeutic decisions broadened. Also, such review could serve an educational function in opening a longer time window on each patient treated by a physician in training. With a database of analyzed narrative covering many patients, the importance of particular features in disease processes could be studied statistically.

Certain recent technological advances favor the development of a computer capability for analyzing narrative (i.e., free-text) input. At the input end, word processors are increasingly used to prepare natural language documents, making it possible to capture narrative data as a byproduct of routine activities. At the same time that the narrative is used for generating reports, it can be input to a processor to create structured information from the free-text input. It was once thought that the amount of memory and computing power that would be necessary to perform "intelligent" text processing were of such magnitude as to preclude the use of these computer systems. Now, one can think realistically of accommodating a natural language processing system in affordable "minis" and even "micros." At the output end, the problem of managing stored data has been aided by an advancing technology of database management systems. Tabular data can be readily input into a database, and increasingly a system user who is not a computer scientist can query a database via a "user friendly interface." These features of the technological environment invite the development of a language processing capability, but they do not create it. For this we need to apply the methods of linguistic analysis to determine both the appropriate information structures for textual information and the algorithms for mapping text into these structures.

LINGUISTIC BASIS OF ALGORITHMS

The basis for natural language information processing is certain observable regularities in the way language is used to convey information, in particular in written texts within specialized technical and scientific domains ("sublanguages") (Harris 1968; Bross et al. 1972; Kittredge and Lehrbenger 1982). In addition to the role of grammatical relations in structuring textual information, it is observed that rules akin to grammatical constraints govern the choice of vocabulary items and their combinations within sentences of a technical specialty. A physician will accept as a medically possible sentence: *Knee hurts*, and reject as not medically well-formed: *Bellevue hurts*, even though both *Bellevue* and *hurts* are within the clinical vocabulary. The sequence *Knee hurts* is grammatically well-formed (except for the dropping of the article), conforming to the elementary sentence type: NOUN + TENSED VERB. As a sequence in the sublanguage of clinical reporting, it consists of words from the sublanguage classes H-PTPART + H-INDIC, (where H stands for Healthcare sublanguage; PTPART stands for Patient Part; and INDIC stands for Indicator of Disease, i.e. signs and symptoms). This is a well-formed word class sequence in this sublanguage. The sequence *Bellevue hurts*, on the other hand, consists of sublanguage word classes H-INST + H-INDIC (where H-INST stands for Healthcare Institutions and personnel). This is not a well-formed sentence sequence in this sublanguage.

By stating the major well-formed sublanguage word class combinations in the syntactic relations found in sublanguage sentences, one composes a sublanguage grammar that can be used by a natural language processor to obtain an informational representation of the content of sublanguage texts (Sager 1978; Grishman and Kittredge 1986). The sublanguage grammar contains the word classes special to the domain, and a set of statement types that are syntactic combinations of these word classes, with specified types of modifiers. The sublanguage statement types, while syntactically arrived at, constitute a set of semantic structures for representing text content. Sublanguage texts are found to consist of sequences of instances of the sublanguage statement types under various connectives (Harris 1988; Harris et al. 1989).

The method of obtaining a domain-specific semantic representation of text content by refined syntactic analysis deserves further discussion. One may approach the problem of semantic representation by defining *a priori* categories and structures based either on expert knowledge of the subject matter or on a general view concerning semantics. In contrast, the sublanguage (syntactic) approach views the specialized use of language in a given domain as the key to the content of the documents. The role of the subject matter expert is to verify and refine the representational structures, but not to invent them. Neither does the linguist invent the representational structures. The linguist applies the same descriptive methods that yield a grammar of a whole language to the texts that employ the restricted vocabulary of the domain. The result is a set of word classes and, in terms of these word classes, a set of syntactic formulas (the sublanguage statement types) that have a semantic interpretation, i.e., that correspond to the different types of information in the sublanguage texts. As a result of the manner by which they were arrived at, the statement types constitute a computable representation. Given the sublanguage grammar, a computer can recognize the occurrence of statement types in sublanguage texts, just as it has been found possible for a computer, given a grammar of a whole language, to recognize instances of the well-formed sentence structures in texts of whole language (computer parsing).

While the methods of sublanguage analysis have been developed in a number of subject areas (in greatest detail in Harris et al. 1989), computer analysis of sublanguage texts is as yet in an early stage of development. The examples drawn on in the remainder of this paper are from the LSP work on clinical narrative (Sager et al. 1987). The system has been adapted for French medical documents (discharge letters) as a part of a joint project with the Division Informatique of the Hôpital Cantonal Universitaire de Genève (HCUG) (Nhan et al. 1989; Sager et al. 1989). The plan is to make narrative processing an integral part of the DIOGENE Hospital Information System, now undergoing a reimplementation as DIOGENE2 after more than 10 years of uninterrupted operation (DIOGENE Staff). Some examples of retrievals from a database of analyzed Lettres de Sortie will be given below.

MEDICAL LANGUAGE PROCESSING

The joint Geneva-New York (HCUG-LSP) project reflects a growing interest within the international medical informatics community in applications of medical language processing. Dujols et al. (1986), after a 3-year experiment with 25,000 texts of patients with respiratory disease, concludes that "the possession of a real medical-administrative database, involving medically indexed text, insures the coherence of all the informations belonging to the same patient and makes their use in logistics and DRG's analysis possible." [DRG = Diagnostic Related Group.] The use of analyzed patient narrative as a source of data for clinical studies has also been discussed (Borst and Scherrer 1988).

An International Working Conference on Computerized Natural Medical Language Processing for Knowledge Representation was held in Geneva in September 1988, sponsored by the International Medical Informatics Association (IMIA), the World Health Organization (WHO) and the Swiss Society for Medical Informatics (SSMI) (Scherrer et al. 1989). In opening the Conference, J.-P. Jardel (of WHO) spoke of the goals of the Conference in relation to the "difficult problem which is central to the future development and usefulness of patient-based information systems: how to translate the 'natural language' of doctors and other health professionals into data which can be managed by a computer." There were 29 presentations of results at this Conference, attesting to the growth of activity on this problem. Another conference, held in Paris in June 1989, on "Informatique et Gestion des Unites de Soins" (Degoulet et al. 1989) devoted a session to Coding Systems and Natural Language Analysis at which the importance of the information in narrative reports for ongoing patient care was stressed.

In the United States, years of experience with medical databases containing exclusively numerical and coded data have convinced some investigators that data capture should include a free-text component. Barnett (1984) states that "the primary dilemma in the use of a computer-based medical record system . . . [is] how to reconcile the physician's custom of recording free-form narrative on a blank page with the computer's need for structure and a pre-defined vocabulary"; he suggests providing "both coded and narrative formats."

A larger context for medical language processing is the entire world of language use in medicine, including the medical literature, indexing vocabularies for the literature (e.g., the National Library of Medicine's Medical Subject Headings, MeSH 1988), coding vocabularies (e.g. the International Classification of Diseases, ICD [2nd edition 1980]), multi-faceted nomenclatures for coding (e.g. SNOMED 1982) and databases of various kinds. The Unified Medical Language System ("UMLS"), a research project of the National Library of Medicine, is designed to facilitate the retrieval and integration of information from machine readable resources of all these types. The first versions of several UMLS components are under construction, including its central vocabulary component, the Metathesaurus (Humphreys and Lindberg 1989).

COMPUTER ANALYSIS OF TEXT

An Example

As an initial illustration of the results of sublanguage text processing, Figure 1 shows a single sentence from a patient document after it has passed through the LSP Medical Language Processor. The tabular display shows in a simplified manner how the analyzed narrative is stored in a relational database. Each sentence has a unique identifier (SID), and each of its component assertions (instances of medical statement types) is assigned a ROW number. The CONJUNCT column contains the linguistic connectives between rows (without showing the scope relations that are stored internally).

The remaining column heads name in a compressed form the different categories of information that may occur in one or another medical statement type. By and large, these categories stand in one-to-one correspondence with the sublanguage word classes. However, the procedure for mapping words into row-column slots does not just match word class with column name. The parsing acts as a control to ensure that the words in a row are a single syntactic unit and do indeed have the informational relation implied by their co-occurrence as instances of the categories represented by their respective column heads.

Was seen in emergency room 2 days ago for diaper rash
and given bacitracin and oral antibiotic.

SID	*ROW*	*CONJUNCT*	*TXTT*	*VERB*	*DIAG_SS_R*	*PR_TM*
01B.01.08	R 01		WAS SEEN IN EMERGENCY_ROOM			[PAST] [P] 2 DAYS AGO
01B.01.08	R 02	"FOR "			DIAPER_RASH	
01B.01.08	R 03	"AND "	WAS GIVEN BACITRACIN			[PAST]
01B.01.08	R 04	"AND "	WAS GIVEN ORAL ANTIBIOTIC			[PAST]

FIGURE 1. Database Table for a Medical Sentence

For example, the heading TXTT stands for column heads (read equivalently: "database fields") covering Test, Examination and Treatment word classes. In Row R01, the word sequence *was seen in emergency room* was mapped into a Treatment field of the type: General Medical Management, based on the subclass H-TTGEN for general management words, here *seen*. (See the next main section below for further details on the dictionary.) The compound noun *emergency room*, treated as a single unit in the dictionary, is in the H-INST class and occurs in a prepositional phrase modifying the H-TTGEN word *seen*; this syntactic combination of medical word classes has been checked for well-formedness during processing.

In R01, the VERB column is empty. The VERB column contains verbs that are not more specifically classified medically, and also verbs, e.g. *show*, that connect words in the TX fields (tests and examination procedures) to words in the DIAG – SS – R fields (Diagnosis, Sign/Symptom, Result). The column PR – TM in Figure 1 combines the fields Precisions and Time, here entirely filled with Time entries. The notation [PAST] comes from the analysis of tense. The [P] in row R01 under PR – TM stands for a (non-occurring) Preposition in the adverbial noun phrase of time: *2 days ago*.

Row R02 contains the Sign/Symptom entry *diaper – rash* (a lexical unit in the subclass H-INDIC). Rows R03 and R04 again contain Treatment entries, here of the type Medication. The words *bacitracin* and *antibiotic* are in the subclass H-TTMED, which determines their mapping into the Medication field; *given* is in the subclass H-TTGEN, which makes a well-formed combination with H-TTMED. Note that the word sequence containing the conjunction *and* in the sentence, namely, *given bacitracin and oral antibiotic*, has been expanded in rows R03 and R04 to contain the implicitly present verb words that were carrying the tense; hence the PR – TM column entries for these rows also contain [PAST].

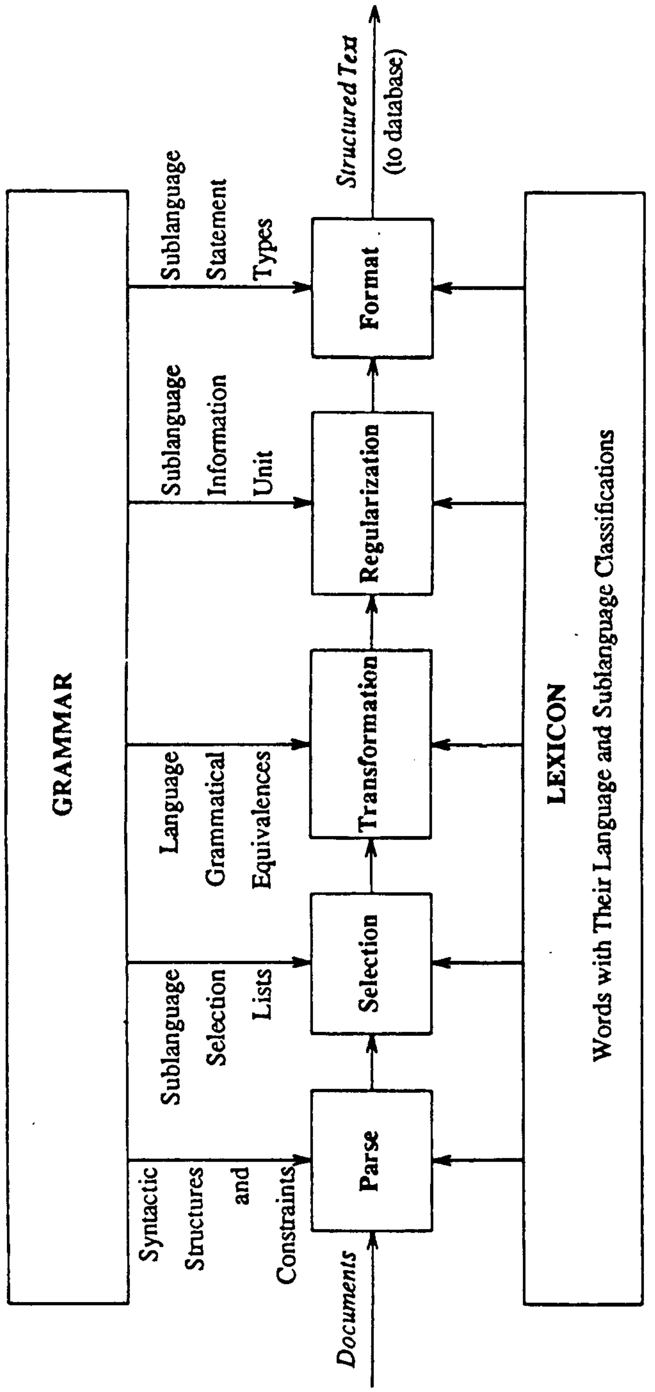
GRAMMAR
Syntactic Structures and Constraints
Sublanguage Selection Lists
Language Grammatical Equivalences
Sublanguage Information Unit
Sublanguage Statement Types
Documents
Parse
Selection
Transformation
Regularization
Format
Structured Text
(to database)
LEXICON
Words with Their Language and Sublanguage Classifications

FIGURE 2. The LSP Medical Language Processor

Stages of Processing

The LSP system is shown schematically in FIGURE 2. It consists of a control program, called "the parser" (not shown in FIGURE 2), that calls on a grammar and a lexicon that contain both whole-language and sublanguage information. The grammar is divided into five successive components. Each component operates on the output of the preceding one. The types of information incorporated in the successive components are indicated in FIGURE 2 by the arrows and captions leading from the GRAMMAR super-box to the individual components. The function of each component will be described by following the example sentence of FIGURE 1 through the five stages of processing.

The first two components, PARSE and SELECTION, together produce a preferred syntactic analysis of the sentence, that is, one that has been "selected" on sublanguage grounds from among alternative correct syntactic analyses. The parser uses a string grammar, in this case, of English (Sager 1981). Two forms of output are available: a tree, and a short form, shown in FIGURE 3. Names of grammatical strings appear to the left of the = sign, and their elements to the right, with the words that satisfy the element written below it. An element may be satisfied by a string (e.g. OBJECT in line 4) in which case the number of the output line on which the string is written appears below the element (e.g. OBJECT in line 4 is satisfied by VENPASS in line 5). If a word has a string as left/right modifier, the number of the output line on which the modifier string is written appears to the left/right of the word.

Note that in line 7 the Noun-STrinG-of-Time (NSTGT) is satisfied by *2 days ago* followed by a modifying prepositional phrase PN on line 10 (*for diaper rash*). This is syntactically correct; a noun (*days*) can be modified by a PN. But in the clinical sublanguage, a PN whose preposition is *for* and whose noun is a symptom word is not a well-formed modifier of a time word, since we do not see occurrences *weeks for fever, days for palpitation*, etc. The preferred reading is to analyze *for diaper rash* as an adjunct of the string in line 5 (as a second SA), or equivalently as an adjunct of the verb *seen*. This checking and moving operation, illustrated by the arrow in FIGURE 3, is the work of the Selection component, operating on the tree produced by the Parse component.

The aim of the third component, Transformation, is to normalize the sentence into ASSERTIONs and FRAGMENTs, each of which will be mapped into one information format, and ultimately into a row of the database representation. The component first recovers the implicit information due to reductions permitted under conjunctions; this is illustrated in FIGURE 4. It also expands relative clause into full assertions, turns (selected) passives into active form and reunifies verbal splits due to past or present participles.

The Regularization component expands the connective structures of the sentence so that they string out in Polish notation format. In Polish notation, the connective (or operator) precedes the two entities it connects (its arguments). This enables scope relations to become explicit without the use of parentheses. The function of the Regularization component with respect to the example sentence is illustrated in FIGURE 5.

```
* CP_01  1B.01.08
* WAS SEEN IN EMERGENCY ROOM 2 DAYS AGO FOR DIAPER RASH
* AND GIVEN BACITRACIN AND ORAL ANTIBIOTIC .

 Parse  1

1.   SENTENCE  =  TEXTLET
                   2.

2.   OLDSENT   =  INTRODUCER  CENTER  ENDMARK
                                3.       .

3.   FRAGMENT  =  SA   TVO    SA
                       4.

4.   TVO       =  TENSE    SA  VERB  SA  OBJECT   SA
                               WAS       5.

5.   VENPASS   =  LVENR  SA   PASSOBJ  SA   ANDSTG
                  SEEN  6.              7.   AND 8.

6.   PN        =  P    NSTGO
                  IN   EMERGENCY_ROOM

7.   NSTGT     =  LTIME   NSTG
                          9. DAYS AGO (10.)

8.   Q-CONJ    =  LVENR  SA   PASSOBJ                 SA
                  GIVEN       BACITRACIN AND 11.

9.   LN        =  TPOS    QPOS   APOS   NPOS
                           2

10.  PN        =  P      NSTGO
                  FOR    DIAPER_RASH

11.  Q-CONJ    =  LN   NVAR          RN
                  12.  ANTIBIOTIC

12.  LN        =  TPOS    QPOS   APOS   NPOS
                                 ORAL
```

FIGURE 3. Parse of Example Sentence

'Was seen in emergency room 2 days ago
for diaper rash
and ___ given bacitracin
and ___ ______ oral antibiotic.'

↓ *Conjunction gap fill-in*

'Was seen in emergency room 2 days ago
for diaper rash
and WAS given bacitracin
and WAS GIVEN oral antibiotic.'

↓ *Subject gap fill-in*

'[] Was seen in emergency room 2 days ago
for diaper rash
and [] WAS given bacitracin
and [] WAS GIVEN oral antibiotic.'

Note: ____ *shows conjunction ellipsis gap,*
[] *shows position of deleted subject,*
Words in capital letters are generated during processing.

FIGURE 4. Example Sentence under Transformation Component

The Format component operates on the regularized parse tree (RPT) produced by the previous component. It is assumed here that each transformed ASSERTION or FRAGMENT in the RPT conforms to a statement type of the sublanguage. Information formats, or FORMATs, are the implemented form of the statement types recognized by sublanguage analysis. The LSP System currently defines three types of format trees: FORMAT1-3 for treatment statement types, FORMAT4 for laboratory statement types and FORMAT5 for patient description as a result of physical examination and history.

At each ASSERTION or FRAGMENT, the component reviews the elements and decides which type of FORMAT fits the phrase. It then links the appropriate RPT nodes with the FORMAT nodes. During this process, modifier nodes, such as TENSE, MODAL, NEGation, QUANTITY, and various TIME nodes are placed in the FORMAT next to their semantic host node. FIGURE 6 shows one form of output from the Format component for the example sentence, a reduced form of the Format Tree. (Note: CTEXT stands

'Was seen in emergency room 2 days ago
for$_1$ diaper rash
and$_2$ WAS given bacitracin
and$_3$ WAS GIVEN oral antibiotic.'

↓ *Connective Preposed*

'and$_2$ for$_1$ [was seen in emergency room 2 days ago]
[diaper rash]
and$_3$ [WAS given bacitracin]
[WAS GIVEN oral antibiotic].'

FIGURE 5. Example Sentence under Regularization Component

for Core text; RTEXT/LTEXT stands for right/left adjunct text, in the case where the adjuncts are not separately formatted.) Another form of output is produced for mapping into a standard relational database management system—such as INGRES or INFORMIX. Each major field in the database table that holds the analyzed documents corresponds to one or more major nodes in the Format Tree. Each Format Tree thus corresponds to one record in the dBMS table.

A DICTIONARY FOR COMPUTERIZED TEXT ANALYSIS

The quality of the lexicon is critical to the success of computerized text analysis. This is especially true on the sublanguage level where the subclass a word belongs to determines which database field the word will occupy, given that all constraints are satisfied. Thus, a crucial part of the sublanguage analysis is to establish the sublanguage word classes and to code the words of incoming texts in terms of these attributes.

FIGURE 7 shows the entries in the LSP English medical dictionary for the words in the example sentence. Under each word are its major categories (parts of speech), separated by commas. Each part of speech may have associated with it a set of attributes, i.e. subclasses that the word belongs to, written in parentheses following a colon. Attributes may in turn have subattributes, written in the same manner.

```
* CP_01   1B.01.08
* WAS SEEN IN EMERGENCY ROOM 2 DAYS AGO FOR DIAPER RASH
* AND GIVEN BACITRACIN AND ORAL ANTIBIOTIC .


(CONNECTIVE (CONJOINED    (CTEXT = 'AND ')))

(CONNECTIVE (RELATION     (CTEXT = 'FOR ')))

(FORMAT1-3   (TREATMENT   (GEN (CTEXT = 'WAS SEEN ')
                               (RTEXT = 'IN EMERGENCY_ROOM ')
                               (EVENT-TIME  (TPREP1 (CTEXT = '[P] '))
                                            (NUM (CTEXT = '2 '))
                                            (UNIT (CTEXT = 'DAYS '))
                                            (TPREP2 (CTEXT = 'AGO ')))
                               (TENSE (CTEXT = '[PAST] ')))))

(FORMAT5     (PSTATE-DATA (S-S  (CTEXT = 'DIAPER_RASH '))))

(CONNECTIVE (CONJOINED    (CTEXT = 'AND ')))

(FORMAT1-3   (TREATMENT   (GEN (CTEXT = 'WAS GIVEN ')
                               (TENSE (CTEXT = '[PAST] ')))
                          (MED (CTEXT = 'BACITRACIN '))))

(FORMAT1-3   (TREATMENT   (GEN (CTEXT = 'WAS GIVEN ')
                               (TENSE (CTEXT = '[PAST] ')))
                          (MED (CTEXT = 'ANTIBIOTIC ')
                               (LTEXT = 'ORAL '))))
```

FIGURE 6. Information Format for Example Sentence

The English grammatical attributes (all those in FIGURE 7 not prefixed with "H-") have been described in detail in Sager (1981) so will only be briefly noted here. The detail required in coding verbs is instructive. A participle (e.g. *given, seen*) carries an attribute OBJLIST that lists the acceptable objects of the verb in the active voice, and a related attribute POBJLIST that lists the acceptable "passive objects" (read: leftovers of the active object after passivization). In either case, if the verb occurs with an adverbial particle DP, or a characteristic preposition P, the particular DP's and P's are listed as values of the respective DPVAL and PVAL attributes of the object string containing the DP or P. Thus, for *given*, with object string DP1, we have *He has given in, given out, given up*; with object string DP2, DP3: *She has given up smoking, She has given it up entirely*. Other OBJLIST values of active *given* include NSTGO (*He has given permission*), NN, NPN and PNN (*We have given her 3 gold injections, We have given instructions to the family, We have given to the caregiver the*

```
AGO               D: (TIMETAG, DRN, H-TMLOC).

AND               SPWORD: (ANDSTG).

ANTIBIOTIC        N: (SINGULAR, COLLECTIVE, NCOUNT1,
                     NONHUMAN, H-TTMED).

BACITRACIN        N: (SINGULAR, H-TTMED).

DAYS              N: (PLURAL, NTIME1, NUNIT, NONHUMAN).

DIAPER_RASH       N: (SINGULAR, H-INDIC).

EMERGENCY_ROOM    N: (SINGULAR, NCOUNT1, H-INST).

FOR               P: ('FOR', H-CONN).

GIVEN             VEN: (OBJLIST:   (DP1: (DPVAL: ('IN', 'OUT', 'UP')),
                                    DP2: (DPVAL: ('UP')),
                                    DP3: (DPVAL: ('UP')),
                                    DP4: (DPVAL: ('UP')),
                                    NSTGO, NN,
                                    NPN: (PVAL: ('TO')),
                                    PNN: (PVAL: ('TO'))),
                        POBJLIST: (DP1: (DPVAL: ('IN', 'OUT', 'UP')),
                                    NSTGO, NULLOBJ,
                                    PN: (PVAL: ('TO'))),
                        H-TTGEN).

IN                P: ('IN'),
                  DP: ('IN'),
                  D: (DLOC3, DRV, DPRED).

ORAL              ADJ: (H-PTPART, H-TTMODE).

SEEN              VEN: (OBJLIST: (NSTGO, SVINGO, SASOBJBE, SVO),
                        POBJLIST: (NULLOBJ VINGO, ASOBJBE, TOVO),
                        VSENT3,
                        H-TTGEN).

WAS               TV: (SINGULAR, PAST, VBE, OBJLIST: (OBJECTBE)).
```

FIGURE 7. Dictionary Entries for Example Sentence

entire dossier). More often one would find these objects in the passive: (*She was given 3 gold injections, Instructions were given [to] the family*). The entry for the verb *was* contains the OBJLIST value OBJECTBE; this is the name of a set of options defined in the grammar for all forms of *be*.

TABLE 1 contains a list of the main medical sublanguage classes currently used by the LSP Medical Language Processor. With minor revisions (and renaming by area), these are the same classes described in Sager et al. (1987) pertaining to work on English clinical narrative. The sublanguage attributes are used not only in the final stage of information formatting and mapping to the database, but also in the course of processing to obtain the correct analysis.

TABLE 1. LSP Medical Word Classes

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 Area * * * * *	
H-TXCLIN	clinical exam. action	*auscultation*
H-TXPROC	exam. procedure	*ultrason, gastroscopie*
H-TXSPEC	test of specimen	*urine analysis*
H-TXVAR	test variable	*glucose, GB, sédiment*
	* * * * * Treatment Area * * * * *	
H-TTGEN	general medical managment	*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 2 formats	*due to, secondaire à*
H-SHOW	V connects test and result	*shows, confirme, montre*

APPLICATIONS

What does a physician hope for from a database of analyzed clinical documents? One example:

> I have a patient who presents the following clinical picture (here a list of signs and associated diagnoses). Are there any patients in the database that look like him? And if yes, what diagnostic procedures did they undergo and what treatment did they receive? And finally, what was their clinical course?

Another example: "Help me with my overload of paperwork. I have patients who qualify for various studies but filling out forms after every encounter is a terrible burden."

Retrieval programs operating on a database of analyzed text could create tables of similar cases and could assume a significant part of the clerical burden. As a demonstration of these possibilities, the Geneva-New York project is developing retrieval programs that operate on the analyzed narrative of 50 Lettres de Sortie of patients in the Digestive Surgery service of HCUG. One application is to fill out automatically a well-established questionnaire concerning patients with acute abdominal pain. This questionnaire was originally developed by de Dombal et al. (1972), then adapted for continental patients in Paris (Association de Recherche en Chirurgie–ARC) (Flamant 1983) and implemented in Geneva (Borst et al. 1987).

One part of the ARC questionnaire concerns the results of gallbladder ultrasound testing (FIGURE 8). The person (or computer) that completes this part of the questionnaire chooses the items that apply to the patient from among the 7 alternatives. We focus on item 3, thickened walls of the gallbladder. Examples of retrieval results for item 3, obtained by an SQL query operating on the INGRES database of analyzed Lettres de Sortie, are shown in TABLE 2 and the corresponding sentences in FIGURE 9.

The logic of the SQL query for determining whether ultrasonography revealed thickened walls ("parois épaissies") of the gallbladder ("vésicule") divides the query into several search criteria. One criterion embodies the medical knowledge that stones in the gallbladder along with signs of infection

1. voies biliaires normales
2. voies biliaires non vues
3. parois vésiculaires épaissies (> 3 mm)
4. lithiase vésicule / voie biliaire principale
5. grosse vésicule (> 10 cm)
6. grosse voie biliaire principale (> 8 mm)
7. dilatation voies biliaires intra/extra-hépatiques

FIGURE 8. Questionnaire ARC (Paris) des douleurs abdominales aiguës: Echographie

TABLE 2. Retrieval for ARC Questionnaire Ultrasound Question, Item 3

POSITIVE RETRIEVALS

sid	row	conj	txproc	verb	sgn_sym	bodypart	quant
16D.1.08	1		ULTRASON ABDOMINAL		LITHIASE	VE1SICULAIRE	
16D.1.08	2	AVEC			E1PAISSIE	VE1SICULE A2 PAROI	MODE1RE1MENT
19B.1.02	1		LES EXAMENS PRATIQUE1S	MONTRENT	PATHOLOGIQUE	UNE VE1SICULE	
20D.1.04	1			SONT	E1PAISSIES	LES PAROIS DE LA VE1SICULE	MODE1RE1MENT
22D.1.03	1		ULTRASON ABDOMINAL			VE1SICULE	
22D.1.03	3	,			E1PAISSIE	PAROI	
40D.1.13	1		US ABDOMINAL		LITHIASE	VE1SICULAIRE	
40D.1.13	2	AVEC			CHOLE1CYSTITE		
71D.1.09	1		ULTRASON ABDOMINAL		CHOLE1CYSTITE LITHIASIQUE		AIGUES
71D.1.09	2	AVEC			E1PAISSIE	PAROI VE1SICULAIRE	
71D.1.09	3	[NMOD]		AYANT	E1PAISSIE	PAROI VE1SICULAIRE	8 MM DE E1PAISSEUR

End of Request - 11 Rows

NEGATED RETRIEVALS

sid	row	conj	txproc	verb	sgn_sym	bodypart	quant	negat
02B.1.01D	2	[NMOD]	UNE ULTRASONOGRAPHIE	METTANT_EN_E1VID	UNE LITHIASE	VE1SICULAIRE		
02B.1.01D	4	[AVEC]			E1PAISSIE	PAROI		[NON]
11D.1.01	1		ULTRASON ABDOMINAL		LITHIASE	VE1SICULAIRE		
11D.1.01	2	[AVEC]			SIGNE INFLAMMATOIRE			[NON]
11D.1.01	3	ET "	ULTRASON ABDOMINAL		LITHIASE	VE1SICULAIRE		
46B.1.03	1		LE_/_LA US	MONTRE	UNE LITHIASE	VE1SICULAIRE		
46B.1.03	2	[AVEC]			SIGNE INFLAMMATOIRE			[NON]
49D.1.05	1		ULTRASON ABDOMINAL			VE1SICULE BILIAIRE	TRE2S GROSSE	
49D.1.05	2	"A2 "			CONTENU LITHIASIQUE			
49D.1.05	3	[AVEC]			SIGNES NETS INFLAMMATOIRES			[NON]

End of Request - 10 Rows

02B.1.01 LE PATIENT SE PLAINT DE DOULEURS E1PIGASTRIQUES DEPUIS PLUSIEURS MOIS , RAISON POUR LAQUELLE VOUS AVEZ PRATIQUE1 DIFFE1RENTS EXAMENS DONT UN TRANSIT OESO-GASTRO-DUODE1NAL MONTRANT UNE PETITE HERNIE HIATALE PAR GLISSEMENT TOUT A2 FAIT RE1DUCTIBLE SANS SIGNES DE REFLUX ET UNE ULTRASONOGRAPHIE METTANT EN E1VIDENCE UNE LITHIASE VE1SICULAIRE AVEC DES VOIES BILIAIRES NON DILATE1ES , SANS PAROI E1PAISSIE .

11D.1.01 ULTRASON ABDOMINAL 86 / 12 / 23 : LITHIASE VE1SICULAIRE SANS SIGNE INFLAMMATOIRE ET SANS DILATATION DES VOIES BILIAIRES .

16D.1.08 ULTRASON ABDOMINAL : LITHIASE VE1SICULAIRE AVEC VE1SICULE A2 PAROI MODE1RE1MENT E1PAISSIE .

19B.1.02 LES EXAMENS PRATIQUE1S MONTRENT UNE VE1SICULE PATHOLOGIQUE : ENTRE POUR CHOLE1CYSTECTOMIE .

20D.1.04 LES PAROIS DE LA VE1SICULE SONT MODE1RE1MENT E1PAISSIES .

22D.1.03 ULTRASON ABDOMINAL 86 / 12 / 28 : VE1SICULE AVEC 1 CALCUL DE 1.6 CM DE DIAME2TRE , PAROI E1PAISSIE .

40D.1.13 US ABDOMINAL DE LE 87 / 03 / 23 : LITHIASE VE1SICULAIRE AVEC CHOLE1CYSTITE .

46B.1.03 LE / LA US MONTRE UNE LITHIASE VE1SICULAIRE SANS SIGNE INFLAMMATOIRE .

49D.1.05 ULTRASON ABDOMINAL DE LE 87 / 04 / 06 : TRE2S GROSSE VE1SICULE BILIAIRE A2 CONTENU LITHIASIQUE , SANS SIGNES INFLAMMATOIRES NETS .

71D.1.09 ULTRASON ABDOMINAL DE LE 87 / 02 / 15 : CHOLE1CYSTITE AIGUE5 LITHIASIQUE AVEC PAROI VE1SICULAIRE E1PAISSIE AYANT 8 MM DE E1PAISSEUR ET OBJECTIVATION DE UN GRAVAT DE PETITS CALCULS DANS LE FOND DE LA VE1SICULE .

1D.1.10 PAS DE DILATATION DES VOIES BILIAIRES .

ccent input: 1=acute, 2=grave, 3=circumflex, 4=cedilla; 5=umlaut.
onventions: L' *becomes* LE/LA; AU *becomes* A2 LE; DU *becomes* DE LA, etc.

IGURE 9. Sentences Corresponding to Ultrasound Retrievals

e.g., with pathology, inflammation, cholecystitis) implies a thickened wall of he gallbladder. This part of the query retrieves as positive results (top of TABLE) 19B.1.02 and 40D.1.13; and as specifically negated findings 11D.1.01, 46B.1.03, nd 49D.1.05. Another criterion is the mention of a thickened wall or walls "parois épaissies") in relation to the gallbladder. This part of the query retrieves n the positive side 16D.1.08, 20D.1.04, 22D.1.03, and 71D.1.09; and on the egated side, 02B.1.01D. The application of any criterion may involve checking everal linguistically connected rows. Further examples of retrievals performed n computer-analyzed narrative are given in Lyman et al. (1989).

CONCLUSION

The paper illustrates how linguistic algorithms based on co-occurrence properties of words in a technical sublanguage can convert the information in sublanguage texts to a structured form. The word classes and statement types of a sublanguage grammar, in conjunction with a computer grammar of the language of origin, have been shown to be effective for mapping textual information into a semantically structured database. As a case in point, narrative patient documents have been computer-analyzed by the Linguistic String Project's Medical Language Processor and retrieval programs written to locate specific types of clinical findings. Retrieval programs can be written to align and compare features of different patients within a disease area to determine their degree of similarity and difference and to complete questionnaires from information originally stated in narrative form. The existence of a computer-analyzed textual database should make it possible to include in studies information that is often omitted due to the cost or difficulty of extracting such information from narrative sources.

SUMMARY

This paper shows how regularities of language usage within a narrow subject area (sublanguage) are used in computerized informational analysis of free-text input. Documents are processed by the NYU Linguistic String Project (LSP) parsing system, which uses a computer grammar of English, a detailedly coded lexicon, English transformations to regularize syntactic structures, and a semantic component based on sublanguage co-occurrence patterns. The workings of the system and an application to free-text medical documents are described. Recent work on French medical documents is included.

ACKNOWLEDGMENTS

The work on French medical documents was supported in part by Grant 3.973.0.87 from the Fonds National Suisse de la Recherche Scientifique. I am indebted to Ngo Thanh Nhan for the preparation of the figures, and to Emile Chi for the database retrievals.

REFERENCES

BARNETT, G. O.
1984 The application of computer-based medical-record systems in ambulatory practice. *New England Journal of Medicine.* **310**:1643–1650.

BORST, F., P. MEYER, F. FELIX, L. BUEHLER, A. JEANJACQUOT AND A. ROHNER
1987 Devenir des douleurs abdominales non specifices. *Medicine et Hygiene* **45**:1910–1913.

Borst, F. and J.-R. Scherrer

1988 Verification of medical knowledge models using free text analysis In *Artificial Intelligence and Cognitive Sciences*, edited by Demongeot, Herve, Rialle, Roche. Manchester: Manchester University Press, pp. 277-283.

Bross, I. D. J., A. Shapiro, and B. B. Anderson

1972 How information is carried in scientific sublanguages. *Science* 176:1303-7.

de Dombal, F. T., D. J. Leaper, and J. R. Staniland

1972 Computer-aided diagnosis of acute abdominal pain. *British Medical Journal* 2: 9-13.

Degoulet, P., J. C. Stephan, A. Venot, and P.-J. Yvon, Eds.

1989 *Information et gestion des unités de soins*, Informatique et Santé 1. Paris: Springer-Verlag.

DIOGENE Staff

The DIOGENE Hospital Information System, Divison Informatique, Hôpital Cantonal Universitaire de Genève, 1211 Genève 4, Switzerland.

Dujols, P., P. Aubas, P. Godard, et al.

1986 Integrated network and medical communication. In *Proceedings of MEDINFO86, the Fifth Conference on Medical Informatics*, edited by R. Salamon, B. Blum, and M. Jorgensen. Amsterdam: Elsevier Science Publishers B. V., North Holland. pp. 348-351.

Flamant, Y.

1983 Aide au diagnostic des douleurs abdominales aiguës, *J. de l'Assoc. de Recherche en Chirurgie (ARC)*, 16 April.

Grishman, R. and R. Kittredge, Eds.

1986 Analyzing language in restricted domains: Sublanguage description and processing. Hillsdale, NJ: Lawrence Erlbaum.

Harris, Z. S.

1968 *Mathematical structures of language*, Section 5.9. New York: Wiley Interscience.

1988 *Language and information*, Bampton Lectures in America 28. New York: Columbia University Press.

Harris, Z., M. Gottfried, T. Ryckman, P. Mattick, A. Daladier, T. N. Harris and S. Harris

1989 *The form of information in science: A test-case in immunology.* Dordrecht: Kluwer Acad. Publ.

Humphreys, B. L. and D. A. B. Lindberg

1989 Building the unified medical language system. In *Proceedings of the 13th Annual Symposium on Computer Applications of Medical Care*. Washington, D.C.: IEEE Computer Society Press.

Kittredge, R., and J. Lehrberger

1982 *Sublanguage: Studies of language in restricted semantic domains*. Berlin: Walter de Gruyter.

Lyman, M. S., N. Sager, E. C. Chi, L. J. Tick, N. T. Nhan, Y. Su, F. Borst, and J. R. Scherrer

1989 Medical language processing for knowledge representation and retrieval *In Proceedings of the 13th Annual Symposium on Computer Applications in Medical Care.* IEEE Washington, D.C.: IEEE Computer Society Press.

Nhan, N. T., N. Sager, M. Lyman, L. J. Tick, F. Borst, and Y. Su

1989 A medical language processor for two indo-European languages. In *Proceedings of the 13th Annual Symposium on Computer Applications of Medical Care*. Washington, D. C.: IEEE Computer Society Press.

Sager, N.

1978 Natural language information formatting: the automatic conversion of texts to a structured data base. In *Advances in Computers*, edited by M.C. Yovits. New York: Academic Press, pp. 89-162.

1981 *Natural language information processing: A computer grammar of English and its applications.* Reading, MA: Addison-Wesley.

Sager, N., C. Friedman, M. S. Lyman, and members of the Linguistic String Project

1987 *Medical language processing: Computer management of narrative data*. Reading, MA: Addison-Wesley.

Sager, N., M. S. Lyman, L. J. Tick, F. Borst, N. T. Nhan, C. Reviellard, Y. Su, and J. R. Scherrer

1989 Adapting a medical language processor from English to French. In *Proceedings of MEDINFO89, the Sixth Conference on Medical Informatics*. Amsterdam: Elsevier Science Publishers B.V., North Holland.

Scherrer, J. R., R. A. Cote, and S. Mandil, Eds.

1989 *Computerized natural medical language processing for knowledge representation*. Amsterdam: North Holland.

1980 *The international classification of diseases: 9th Revision, Clinical Modification: ICD-9-CM*, 2nd ed. Washington D.C.; U.S. Health Care Financing Administration.

1982 *SNOMED systematized nomenclature of medicine*. Chicago: College of American Pathologists.

1988 *Medical subject headings*. U.S. National Library of Medicine Publications.