# A CODASYL-TYPE SCHEMA FOR NATURAL LANGUAGE MEDICAL RECORDS

N. Sager,* L. Tick,‡ G. Story,* L. Hirschman*

*Linguistic String Project, Faculty of Arts and Science
‡School of Medicine, NYU Medical Center
New York University, New York, New York

Summary - This paper describes a CODASYL (network) database schema for information derived from narrative clinical reports. The goal of this work is to create an automated process that accepts natural language documents as input and maps this information into a database of a type managed by existing database management systems. The schema described here represents the medical events and facts identified through the natural language processing. This processing decomposes each narrative into a set of elementary assertions, represented as MEDFACT records in the database. Each assertion in turn consists of a subject and a predicate classed according to a limited number of medical event types, e.g., signs/symptoms, laboratory tests, etc. The subject and predicate are represented by EVENT records which are owned by the MEDFACT record associated with the assertion. The CODASYL-type network structure was found to be suitable for expressing most of the relations needed to represent the natural language information. However, special mechanisms were developed for storing the time relations between EVENT records and for recording connections (such as causality) between certain MEDFACT records. This schema has been implemented using the UNIVAC DMS-1100 DBMS.

## Introduction

The large bulk of clinical information, aside from laboratory data, is recorded in free text form. This paper addresses the subject of the construction and use of a data base for information derived from freely written narrative describing the events related to a patient's state of health and care, such as the history and course of an illness. If computer programs could organize and access such information, this would create a valuable supplement to the numerical and codified questionnaire-derived databases currently being developed in a number of disease areas. Qualitative, nuanced observations and the time course of events could then be queried along with the "hard data" of the medical condition.

The techniques for creating databases that include information from narrative sources exist. We have shown in previous work that the language used to report clinical findings is sufficiently constrained both by the nature of the subject matter and by stylized habits of usage, so that it has been possible to construct language processing programs that convert the information from its free-text form to an equivalent table of information.[1,2] In this table, called an information format, the time relations of events reported in the documents are given an explicit representation[3] and instances of each different type of clinical finding (sign/symptom, diagnosis, therapy, etc.) are grouped together under an appropriate heading. When the narrative information is structured in this way, retrieval programs can determine whether certain events preceded other events, whether given procedures were carried out and with what effect, etc.[4,5]

The creation of a data base from clinical narrative can be viewed as consisting of two main steps: (1) processing the language material to obtain a structured form of the information; (2) mapping the structured information into the form needed for the storage and selective retrieval of the information, i.e., into a system for managing the data base. This system should have certain desirable characteristics which would allow accessing programs to be relatively ignorant of the storage arrangements, maintain some sort of access control, and provide a degree of data integrity as needed; in short, all those good features of what has come to be known as a data base management system (DBMS).

Since in this case the objects to be "managed," i.e., the analyzed clinical narrative, are, or certainly appear to be, different from those which appear in the DBMS literature, one approach is to undertake de novo the design of a DBMS for natural language data bases which incorporates the desirable features mentioned above. Though this seems attractive as a direct approach to the data handling problem, such efforts can get out of hand. In the June 1980 issue of the Transactions on Database Systems, there appears a paper by **Stonebraker** describing a project whose goal was to produce a usable (relational) DBMS. Even a casual reading of the **Stonebraker** paper shows that a large portion of the human effort, and likely of machine time, went into what might be described as a small portion of the intellectual work concerned with the data base subject proper, e.g., data organization and accessing mechanisms. Utilities, security, concurrency, documentation, training, etc., consumed most of the project's resources.*

In view of these considerations, in our own work it was decided to examine existing data models and their associated DBMS's as to their appropriateness

*The aforementioned paper is especially recommended to those who retain any belief in the notion that a big system is a bigger small system. The pain on each page was all but literally transmitted to one (LJT) of the present authors.

0195-4210/80/0000-1027$00.75 © 1980 IEEE

for managing structured narrative data. In this paper we are then primarily concerned with the processes for mapping the outputs of the medical language processor into the structure required by an existing DBMS-data model. (Here, "existing DBMS" is to be understood as including systems which will be available within the next five years or so.) In addition to the efficiency of development gained by working with existing data models, a model provides a formalism against which to test and refine the products of the language analysis with regard to their informational efficacy and consistency.

To a much larger extent than predicted, it was possible to cast the analyzed narrative into the formalism of a CODASYL data base schema. After a brief summary of the language processing, the remainder of the paper will describe this schema and some of the design decisions that were made. Not all features of the narrative information fit the model; the connective structure of sentences that gives language its distinctive ability to carry reasoning and other higher level connections among subsidiary statements had to be represented as data components rather than relationships embedded in the schema; temporal relations among reported events also required a special representation.

## Language Processing

The medical language processor developed by the Linguistic String Project of New York University, as it presently operates, converts the information in clinical narrative into a structured form by a four-step process. The input sentences are first analyzed syntactically using a parsing program[7] equipped with a broad-coverage English grammar[8] which has been augmented and specialized for medical usage. The parsed sentences then undergo regularization via information-preserving transformations that reduce the number of different syntactic structures in the parse trees; e.g., the passive form is changed to the active form, conjunctional constructions and relative clauses are expanded (e.g., patient had stiff neck and fever → patient had stiff neck and patient had fever), etc. The result of these two steps, parsing and regularization, is a decomposition of each sentence into component elementary assertions (more specifically, assertion parse trees) which may be connected in a binary fashion by linguistic connectors.

The third step of language processing, called information formatting, draws on informational classifications of the medically-specific vocabulary in a process that rearranges and relabels the regularized parse trees so as to display the types of medical statements that are present. It has been shown that the English sentences and phrases used in medical reporting fall into a small number of types, different combinations of a limited number of informationally significant word classes.[9] The same is true in many disciplines, so that one can speak of the language used as the "sublanguage" of the discipline. An information format is one way of representing successive statements as instances of sublanguage word-class patterns, or sublanguage statement types, that are characteristic of discourse in the discipline.

The fourth step in the conversion of free narrative to structured information is "normalization," a process whereby the computer program examines the information-formatted sentences in the neighborhood of a given one, in order to fill in missing information which can be deduced from context. In the case of patient records, document heading information (e.g., dates of admission and discharge in a hospital discharge summary) and paragraph headings may also be consulted. The main types of implicit information supplied in the normalization of formatted clinical narrative are part-of-body information (e.g., throat supplied for hoarseness) and time information.

## MED-RECORDS Schema

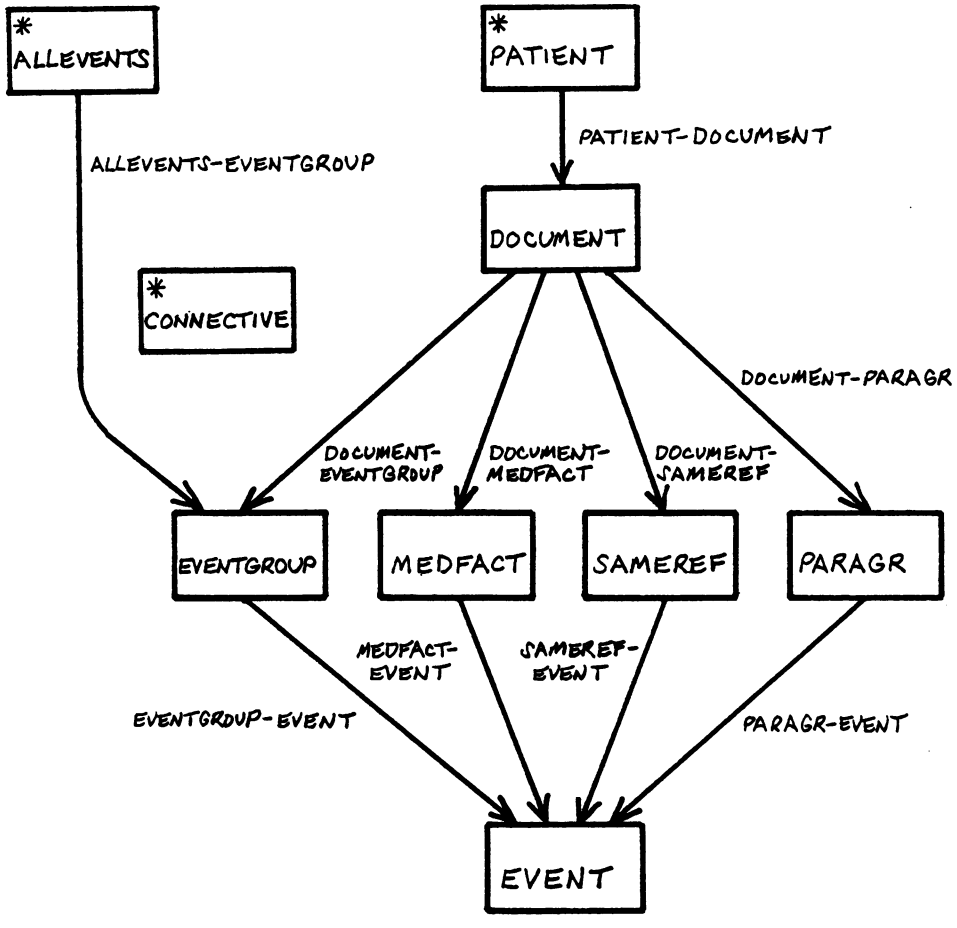
The CODASYL-type[10] schema described here reflects the underlying linguistic structure of the text. A number of sample narrative discharge summaries were used in the design of the schema, called MED-RECORDS; a diagram of the schema is shown in Figure 1. The boxes indicate the different record types (such as EVENT), and the arrows indicate CODASYL "set" relationships. A CODASYL set is a relationship in which an instance of the owner record type is said to "own" instances of the member record type. Thus in the MEDFACT-EVENT set a single MEDFACT record instance "owns" several EVENT record instances, as discussed below.

### MEDFACT Record

Each MEDFACT record instance corresponds to a single elementary assertion from a sentence of the input document. (Each document sentence contains one or more elementary assertions.) The linguistic analysis has broken each assertion into medical events corresponding roughly to subject and predicate; these can be classified into a small set of informational categories (see top of Figure 2). The record type EVENT is used to hold a medical event; a MEDFACT record instance owns (via the MEDFACT-EVENT set) the EVENT records which correspond to the subject and predicate of the assertion. The various information categories for the medical events are shown in Figure 2; the EVENT-TYPE data item in the EVENT record holds a value which records the category of information.

Each of the medical events from a single elementary assertion has a different EVENT-TYPE value; the input texts processed thus far indicate that these values occur in a limited number of combinations. Both the MEDFACT and EVENT records contain a data item, MEDFACT-TYPE, which gives the particular combination of information categories involved in a given assertion, i.e., in an instance of the MEDFACT-EVENT set (see Figure 2).

Each elementary assertion in the natural language text may contain various modifiers: evidential, aspectual, time, etc. An issue in the schema design has been to what extent these modifiers can be moved from the assertion level (their linguistic home) down to the level of the medical events, i.e., from the MEDFACT record to the EVENT records. Placing modifiers in the EVENT record makes it possible in some retrievals to access only the EVENT record; the MEDFACT record may contain no extra information relevant to the retrieval and could be ignored in these cases.




Boxes indicate record types.

Arrows indicate set relationships (owner → member).

The set name appears adjacent to the arrow indicating the set.

"*" indicates location mode direct, calc, or index sequential.

The CONNECTIVE record is linked to other records by pointers stored within the record rather than by participation in a set.

Figure 1. The CODASYL-type schema MED-RECORDS.

EVENT-TYPE values

| | | |
|---|---|---|
| 01 | V-MD | medical administrative actions<br>This is the fourth admission. |
| 02 | V-TR/RX | medicine and treatments<br>She was given folic acid 2.5 mg PO daily. |
| 03 | LAB | lab tests<br>Urine - urinalysis normal. |
| 04 | DEVEL | patient growth and development<br>...well developed, well nourished. |
| 05 | EXAM-TEST | tests and techniques used during a physical examination<br>Liver palpable, 4 cm; spleen -- 2 cm. |
| 06 | BODY-MEAS | routine body measurements<br>TMP 102, PU 175, RR 75, weight 15 lbs. |
| 07 | BODY-FUNC | description of body functions<br>Appetite was good. |
| 08 | LAB-RES | qualitative results of lab tests and cultures<br>Agent was h. influenzae. |
| 09 | S-S | signs or symptoms<br>Throat was hyperemic, no exudate. |
| 10 | DIAG | diagnoses<br>...female, who is known to have sickle cell disease. |
| 11 | DESCR | neutral descriptors<br>...patient was comfortable, eating well and receiving maintenance IV fluids only. |

MEDFACT-TYPE values

| | | |
|---|---|---|
| 01-11 | | same as EVENT-TYPE above (these correspond to MEDFACT records which own only one EVENT record) |
| 12 | V-MD + V-TR/RX | She was admitted for pleural biopsy. |
| 13 | V-MD + S-S | This is the 2nd admission of ... for fever of 2 days. |
| 14 | V-MD + DIAG | This is the 1st hospitalization of ... with septicemia. |
| 15 | V-MD + LAB | Admitted for tomography. |
| 16 | LAB + LAB-RES | Blood cultures positive for pneumococcus. |
| 17 | LAB + S-S | Chest x-rays suggested pulmonary infarct ... |
| 18 | LAB + DIAG | Chest x-rays revealed progressive bilateral pneumonia. |
| 19 | EXAM-TEST + S-S | Pain on dorsiflexion. |
| 20 | BODY-FUNC + S-S | Loss of appetite. Difficulty in breathing. |
| 21 | BODY-MEAS + S-S | Allorhythmic pulse. |
| 22 | DEVEL + S-S | Arrested development. |

Figure 2. Values for data items EVENT-TYPE and MEDFACT-TYPE.

The original text sentence from which a MEDFACT record instance was derived may be accessed via a pointer in the MEDFACT record. The entire paragraph can be retrieved through the PARAGR record (see below).

EVENT Record

Each instance of an EVENT record corresponds to a single mention in the document of a medically significant event. Figure 3 shows the data items of the EVENT record in detail, with the appropriate items filled for the text words, liver palpable, 4 cm. The data item EVENT-TYPE gives the informational category into which the event falls; in this example palpable is coded 05 EXAM-TEST. The actual text of the medical event to which the EVENT record corresponds is found in the data item EVENT-TEXT. (All the data items which hold text have subfields for adjuncts which occur to the left and right of the text in the original sentence.) Because so much of the data involves particular body parts, the EVENT record contains a data item for this information, should it be present in the assertion. In the example of Figure 3, the BODYPART data item has the value "liver."

| | |
|---|---|
| EVENT-TYPE | 05 |
| EVENT-TEXT | palpable |
| MEDFACT-TYPE | 05 |
| AUX-INFO | -- |
| BODY-PART | liver |
| NORMALCY | -- |
| NEGATION | -- |
| MODALITY | -- |
| CHANGE | -- |
| ASP-TYPE | -- |
| ASP | -- |
| REPT | -- |
| INST | -- |
| NUMQ | 4 cm |
| NONNUMQ | -- |
| VERB | -- |
| GEN-REF-PT | 03 |
| BEG-TIME | 14 |
| END-TIME | -- |

Figure 3. The EVENT record in detail, with source text "Liver palpable, 4 cm."

The EVENT record contains a number of data items for modifiers which have been moved down from the MEDFACT "level." NEGATION contains words of negation, as in "no fever." ASP contains aspectual words; ASP-TYPE indicates one of three categories: beginning, ending, or durational. MODALITY contains words which express uncertainty, as in "pain seemed to subside."

Other data items in the EVENT record have been determined through analysis of the sample discharge summaries. The data item NORMALCY holds words indicating normalcy or a change toward normalcy, such as "breathing normally" or "chest x-ray negative." The CHANGE item holds words describing change ("appetite improved"). The REPT item indicates repetition of an event, e.g., "2 x-rays," or a plural marker as in "seizures." Events are sometimes quantified either numerically or non-numerically (severe, mild). This information is placed in the NUMQ and NONNUMQ data items, respectively.

Occasionally an event is described which occurred or will occur at an institution different from that which supplied the document. In that case the name of the institution is put into the data item INST. The assertions of the sample data often contain verbs which carry little information themselves and which usually serve to connect the more important medical events. Examples are "patient has meningitis" and "blood culture was negative." Though it is expected that these words will have little importance in retrieval, they are retained in the data item VERB in the EVENT record.

A fundamental characteristic of narrative is that the events reported occur in time and there is an implicit forward progression of time as the narrative proceeds (provided that no explicit time information is given). In the processing stage called normalization (see above) the time-points of the narrative are determined. Each medical event, and thus each EVENT record instance, is assigned a beginning time point and, for events of duration, an ending time-point. The data items BEG-TIME and END-TIME in the EVENT record correspond to these time-points. The data item values are indices in a time-point matrix stored in the DOCUMENT record. A fuller discussion of the representation of time information is found in the description of the DOCUMENT record below.

The dates of admission and discharge are distinguished time-points for medical records describing a hospital stay, and they allow the events reported in the document to be divided into five general time-periods: (1) prior to admission, (2) at admission, (3) during hospitalization, (4) at discharge, and (5) after discharge. It is convenient, when programming retrievals requiring only gross time constraints, to have available in the EVENT record instance the particular time-period during which the event occurred. This information is found in the data item GEN-REF-PT (general reference point).

### EVENTGROUP and ALLEVENTS records

The EVENT records are naturally partitioned by the information categories to which they belong (EVENT-TYPE). For each possible value of the data item EVENT-TYPE there is an EVENTGROUP record instance which owns (via the EVENTGROUP-EVENT set) all the EVENT records with that EVENT-TYPE from a given document (see Figure 1). Such a partition of all the EVENT records in the entire database is given indirectly via the ALLEVENTS record. Each ALLEVENTS record instance owns all the EVENTGROUP records throughout the database which correspond to a single EVENT-TYPE value.

### CONNECTIVE Record

As has been stated above, each sentence of the document is decomposed into one or more elementary assertions, represented in the schema by instances of the MEDFACT-EVENT set (each MEDFACT record owns EVENT records). Assertions (and thus MEDFACT record instances) from the same sentence are joined in a binary tree structure by conjunctions and other linguistic connectives. Examples of such trees are shown in Figure 4. These structures group the MEDFACT records on a sentence by sentence basis, and

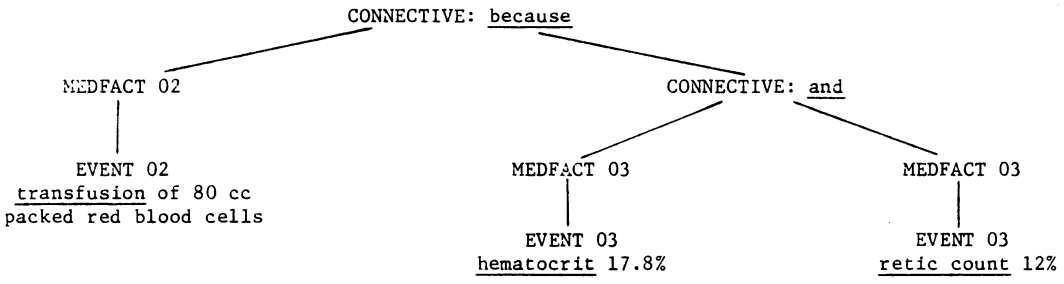



Figure 4. CONNECTIVE/MEDFACT binary tree for text "Transfusion of 80 packed red blood cells given because HCT was 17.8% and retic count was 12%."

the connectives can carry important information, such as a causal relationship between events. The CONNECTIVE record holds the connective text; the CONN-TYPE data item contains a code identifying the connective type. The set mechanism in the current CODASYL specification was found unsuitable for representing the binary tree structures; they are implemented instead by pointers stored in the CONNECTIVE and MEDFACT records. Since this structuring is not seen in the schema itself, navigation of the binary trees must be programmed by the user.

The issues addressed in the MED-RECORDS schema design fall roughly into two categories: (1) representation of the information found within a sentence, and (2) representation of the information given by intersentential relationships. The schema aspects discussed above (primarily the EVENT, MEDFACT, and CONNECTIVE records and the sets relating them) involve the first category of issues addressed, that of "intrasentential" information. The second category of information deals with those elements which characterize a coherent narrative: implicit progression of time, reference, and implicit commonality of topic. This information, which is captured in the language processing step called "normalization," involves the DOCUMENT, SAMEREF, and EVENT records and the sets which connect them.

Time Relationships

Among the relationships which characterize a coherent narrative, those having to do with time are particularly important in documents which report events in a medical care setting. For each document the processing stage called normalization produces (1) a set of time-points, (2) a partial ordering of these time-points over time, and (3) an identification of each medical event with at least one of the time-points. The procedure makes use of both the implicit forward progression of time in narrative and the specific linguistic mechanisms which express time in the input text. Each of the time-points generated by normalization is given a unique number. The transitive closure of the partial ordering is computed, and this result is represented as a bit matrix in the DOCUMENT record. Given any time-points i and j generated by normalization, i precedes j in time only if the bit is set in position (i,j) of the bit matrix. If the length of time interval (i,j) is known, that value can be found in a hash table also stored in the DOCUMENT record. As mentioned above, each EVENT record instance has data items BEG-TIME and END-TIME, corresponding to the beginning time-point and, for durational events, the ending time-point, of the medical event given by the record. The values found in these data items are row and column numbers of the bit matrix in the DOCUMENT record. Thus the DOCUMENT record must be accessed in order to perform time comparisons. The programs for producing and manipulating the bit matrix are still under development. Issues of time-point representation that have not been addressed include the fuzziness of time-points, multiple time-points for a single event (e.g., "patient was given aspirin every four hours"), and the relative ordering of time-points from different documents.

SAMEREF Record

In a natural language narrative report there are often multiple references to the same medical event, achieved by the use of such linguistic entities as pronouns, articles, and relative clauses. In the sample documents used in this research, there are also duplicate mentions of the same event found in different paragraphs (e.g., fever mentioned under "Admission Symptoms" and "patient had fever on admission" under "History of Present Illness"). During the linguistic processing these multiple references are to a large degree identified by (1) resolving antecedents of referentials, (2) identifying the simultaneity of events through their time-points, and (3) noting identity of reference in repetitions of words supplied during English regularization. Separate EVENT record instances which correspond to multiple references to the same medical event are grouped via the SAMEREF-EVENT set, in which a single SAMEREF record instance owns several EVENT records. This serves to collect all the information about a particular medical event and facilitate a check that separate references to an event are not taken as separate occurrences.

The SAMEREF record, by collecting multiple references, brings the schema closer to a model of the actual events which occurred in the hospital, rather than the document which reports those events. Nonetheless the document, with its narrative discourse and particular paragraph structure, is

reflected in the schema. The PARAGR record serves to group the EVENT records (via the PARAGR-EVENT set) according to paragraph, and the order in which the EVENT records are stored in this set is the order in which their corresponding medical events occurred in the text. As was mentioned above, the binary trees formed from the CONNECTIVE and MEDFACT records preserve the sentence boundaries. It is thought that some of the normalization procedure might be implemented using the DBMS, in which case the schema elements which supply the sentence context information would be essential. For some of the linguistic elements it is difficult to decide a priori whether or not they are to be useful in retrieval; the philosophy has been to retain the information (linguistic or otherwise) which is captured, even if its relevance is not immediately recognized.

### PATIENT and DOCUMENT records

Little has been said concerning the PATIENT and DOCUMENT records. Each PATIENT record instance owns the DOCUMENT record instances which correspond to the documents found in the patient's medical history. Both records contain conventional fixed-field information, such as patient name and birth date, document source and identification number, admission and discharge dates for the hospitalization, etc. The one exception is the time information, discussed earlier, found in the DOCUMENT record. Design questions which concern the organization of medical records by an institution are not germane to the issues addressed by this research.

## Conclusion

As a first step toward the possible use of current DBMS technology for the management of a database of analyzed free narrative medical records, a CODASYL schema was designed. The goal was to embody in the database structure the entities and relations captured in the narrative processing. Most of the relationships were mapped into the schema itself and the remainder (specifically, sentence connectors) were embedded in data items.

This schema was tested using the UNIVAC DMS-1100 (CODASYL) DBMS. At the time of writing, some very simple storage and retrieval operations have been performed, and the loading of a modest amount of analyzed narrative is in progress. This is to be followed by making queries of the database using the available DMS 1100 query language.

The goal of this research is two-fold. First to develop guides for the mapping of analyzed narrative into a network schema, and second, perhaps more important, to indicate shortcomings in current design of DBMS's to be used for narrative database management. The ability to use current DBMS's for analyzed medical narrative will allow relatively easy access to qualitative clinical information. In addition, a "standard" DBMS brings with it the potential use of high level query languages to facilitate retrieval and manipulation of the stored clinical data.

## References


1. Sager, N., Natural Language Information Formatting: The Automatic Conversion of Texts to a Structured Data Base. In Advances in Computers 17 (M. C. Yovits, ed.), 89-162. Academic Press, New York, 1978.

2. Sager, N., Hirschman, L. and M. Lyman, Computerized Language Processing for Multiple Use of Narrative Discharge Summaries. In Proceedings of the Second Annual Symposium on Computer Applications in Medical Care (F. H. Orthner, ed.), 330-343. IEEE, New York, 1978.

3. Hirschman, L., Retrieving Time Information from Natural Language Texts. Proceedings of the Joint British Computer Society and Association for Computing Machinery Symposium: Research in Information Retrieval, in press.

4. Sager, N. and M. Lyman, Computerized Language Processing: Implications for Health Care Evaluation. Medical Record News 49, 3 (June 1978), 20-30.

5. Hirschman, L., Sager, N. and M. Lyman, Automatic Application of Health Care Criteria to Narrative Patient Records. In Proceedings of the Third Annual Symposium on Computer Applications in Medical Care (R. A. Dunn, ed.), 105-113. IEEE, New York, 1979.

6. **Stonebraker, M., Restrospection on a Database System. Transactions on Database Systems 5, 2 (1980), 225-240.**

7. Grishman, R., Sager, N., Raze, C. and B. Bookchin, The Linguistic String Parser. AFIPS Conference Proceedings 42, 427-434. AFIPS Press, Montvale, New Jersey, 1973.

8. Sager, N., Natural Language Information Processing: A Computer Grammar of English and Its Applications. Addison-Wesley, Reading, Mass., in press.

9. Hirschman, L. and N. Sager, Automatic Information Formatting of a Medical Sublanguage. Sublanguage: Studies of Language in Restricted Semantic Domains (R. Kittredge and J. Lehrberger, eds.). Series on Foundations of Communication (R. Posner, ed.), Walter de Guyter, Berlin, in press.

10. CODASYL Data Base Task Group Report. Association for Computing Machinery, New York, New York, 1971.


## Acknowledgments


**This research was supported in part by National Library of Medicine grant number LM-02616, awarded by the National Institutes of Health, DHEW; and in part by the National Science Foundation under grant number IST-7920788 from the Division of Information Science and Technology.**