# Sublanguage

## Studies of Language in Restricted Semantic Domains

edited by
Richard Kittredge and John Lehrberger

Walter de Gruyter · Berlin · New York
1982

Foundations of Communication
Library Edition

Editor
Roland Posner


Satz und Druck: Arthur Collignon GmbH, Berlin 30
Buchbinder: Lüderitz & Bauer, Berlin

Chapter 2

# Automatic Information Formatting of a Medical Sublanguage

Lynette Hirschman and Naomi Sager

## 1. *Introduction*

In this paper we are concerned with computer methods for structuring and extracting the information in a body of natural language texts written in a *sublanguage*. We define sublanguage here as the particular language used in a body of texts dealing with a circumscribed subject area (often reports or articles on a technical speciality or science subfield), in which the authors of the documents share a common vocabulary and common habits of word usage. As a result, the documents display recurrent patterns of word co-occurrence that characterize discourse in this area and justify the term sublanguage.

The characteristic word co-occurrence patterns of the sublanguage are central to processing sublanguage texts. This regularity makes it possible to determine a set of sublanguage-specific word classes which correlate with the types of information conveyed in the subfield; these word classes form the bridge between the structure (syntax) of the sublanguage texts and their informational content (semantics). We use the co–occurrence patterns to design a representation for the information in the subfield texts that is adequate for further computer manipulation (e. g., a representation in tabular form). An automatic procedure then converts the free text into the chosen representation, based on membership of the sentence words in the sublanguage classes. Successful processing is highly dependent on the existence of these subclasses and on the ability of the processor to recognize occurrences of these classes in particular syntactic relations.

One of the first investigations into the structure of a science sublanguage was undertaken in 1969, under sponsorship of the National Library of Medicine, with textual material drawn from the pharmacology literature. This work was based on the hypothesis that relevant catgories for representing information in texts of a medical subfield could be obtained by the application of formal (in principle, computable) grammatical methods of analysis to the texts of the sublanguage. From the point of view of linguistics, this was a test of the correlation of word distributional patterns with information; from the point of view of information science, this was a new method for obtaining a characterization of document content.

The results of this initial study are presented in an article reprinted in this volume [Sager 1972]. The study showed that it was possible to write a sublanguage-specific grammar covering texts in a narrow science subfield

and that the word classes established on distributional grounds did indeed correlate with the classes of objects and the relations of special interest in the subfield. A subsequent computer experiment verified this hypothesis: a clustering program was written to operate on syntactically analyzed sentences of subfield texts [Hirschman et al. 1975]. It grouped together words occurring in a similar syntactic and lexical environment (e. g., nouns occurring as the object of the same verb, or verbs occurring with the same subject noun). Words that occurred in a number of similar environments were grouped together into a class. The word classes obtained by the clustering program were semantically coherent and correlated well with those obtained by manual analysis.

The initial study also put forward the concept of an information format as a representation for the information in sublanguage texts. The information format is a table-like structure whose columns correspond to the major sublanguage word classes. Different combinations of the columns correspond to the basic sublanguage sentence types (the characteristic combinations of sublanguage word classes in the SVO (subject-verb-object) relation). For example, in the sublanguage of clinical reporting, a frequent SVO sequence consists of a subject from the PATIENT class (usually *patient* or a pronoun) followed by a VERB in the V-PT class (a verb whose subject is characteristically a PATIENT noun in the sublanguage, e. g., *have, develop*), followed by a word in the SIGN-SYMPTOM class (e. g., *cough, dyspnea*). This basic sentence type (SVO = PATIENT + V-PT + SIGN-SYMPTOM) is seen in such text occurrences as *Patient has had a cough for the past 4–5 months* (SVO = *patient + have + cough*) and *She had an episode of severe dyspnea while in Maryland* (*SVO = she + have + dyspnea*)[1]. The information format organizes the sublanguage sentence types into a compact tabular representation so that the document content can be quickly inspected. When the procedure for mapping sublanguage texts into their format representation is invoked, its output is a set of structured documents which can serve as a data-base for automated fact retrieval and for other data processing operations which could not be performed on the (unprocessed) free narrative [Sager 1978].

In order to represent the information uniformly, the syntactically conveyed connections are translated into the occurrence of particular combinations of column entries in the format. The information-formatting process relies on:

1. a lexicon which provides lexical information about a word's English class, e. g., *cold* = NOUN (with English subclasses count-noun, non-human), and ADJECTIVE; in addition, the lexicon provides information about a word's sublanguage classes, e. g., in the clinical reporting sublanguage, *cold* has as sublanguage classes SIGN-SYMPTOM for the NOUN, and DESCRIPTOR for the ADJECTIVE.
2. a parse, which specifies subject-verb-object relations, host-modifier relations, etc. For example, in *The patient has no cold or fever*, the parse should show that both *cold* and *fever* are nouns, modified by the article *no*, occurring as the object of the verb *has*; the parse should exclude the analysis of *cold* as an adjective, as well as the analysis of *no* modifying *cold* only.
3. a set of paraphrastic transformational relations which reduce complex syntactic structures to informationally equivalent simple structures. In the process, they fill in material omitted due to ellipsis. For example, *The patient has no cold or fever* can be transformed into two complete simple assertions, *The patient has no cold and the patient has no fever.*
4. a mapping of the parsed, transformed structure into the appropriate columns of the information table or format, as Fig. 1 shows in simplified form.

| | PATIENT | PT-STATUS | | |
|---|---|---|---|---|
| | | V-PT | NEG | SIGN-SYMPTOM |
| 1. | patient | have | no | cold |
| | patient | have | no | fever |
| 2. | patient | be | | sick |

Sentence 1: Patient has no cold or fever.
Sentence 2: Patient is sick.

Fig. 1: Partial information format with two sample formatted sentences.

The next section will describe the process of determining the sublanguage classes and designing the information format for a sublanguage of hospital discharge summaries. This will be followed by a section that describes and illustrates the automatic parsing, transforming and formatting of the sublanguage sentences. The final sections will discuss applications of the information formatting technique and particular problem areas that are the subject of on-going research.

## 2. *Sublanguage Analysis*

### 2.1. Word Class Formation

The sublanguage of clinical reporting, as evidenced in a typical patient document, conveys information about the patient's state and medical

[1] The SVO sentence type is defined after a transformational analysis has been performed on the sentence (see section 3.2). Transformations extract the tense and aspectual qualifiers like *episode of, period of*, etc. Hence SVO = *she + have + dyspnea* rather than SVO = *she + have + episode* in the above example.

actions taken in connection with that state. The word classes for this material were developed by examining and comparing sets of words occurring in particular syntactic environments. To illustrate this process, we begin by looking at the noun phrases that occur as the object of the verb *develop* with the subject *patient*, in a sample corpus of eight hospital discharge summaries.

The noun phrases appearing in this environment (shown in Fig. 2) name signs or symptoms of abnormal conditions: *mild cold, fever, severe pain in abdomen, cough, severe respiratory distress.* These noun phrases consist of a head noun which is a sign or symptom word (e. g. *cold, pain, distress*, etc.) with optional modifiers. The environment *patient develop ( )* thus defines a set of words (the heads of the object noun phrases) describing a sign or symptom of an abnormal condition (called here the SIGN-SYMPTOM class).

| SUBJECT | VERB | OBJECT |
|---|---|---|
| patient | develop | mild cold |
| patient | develop | fever |
| patient | develop | severe pain in abdomen |
| patient | develop | cough |
| patient | develop | severe respiratory distress |

Fig. 2: Object noun phrases of the verb *develop* with subject *patient*.

Next, to extend this SIGN-SYMPTOM class, we look for other subject-verb-object combinations which take an object from this class. For example, in the same corpus, we find *fever* occurring in the environment *patient have ( )* in the sentence *2.5 weeks prior to admission, patient had fever and anorexia.* Thus *patient have ( )* is a candidate environment for occurrences of the SIGN-SYMPTOM class. And indeed we also find *anorexia* and *allergies* in this environment, shown in the first two examples of Fig. 3.

Note that the negation of the sign or symptom has the same distribution as the assertion of the sign or symptom, e. g., *patient has no known allergies* and *patient has allergies.* The verbs *complain* and *lapse* also fit into this pattern of distribution, e. g., *On the evening before admission patient complained of stomach ache.*, and *After completion of 4 hour transfusion, patient lapsed into coma.*

On the basis of these occurrences, we can now define a verb class whose subject is *patient* and whose object is a word from the SIGN-SYMPTOM class. This class (called the V-PT class) includes *develop, have* (in certain of its occurrences), *complain of*, and *lapse.*



| | SUBJECT | VERB | OBJECT |
|---|---|---|---|
| 1. | patient | have | no known allergies |
| 2. | patient | have | anorexia |
| 3. | patient | have | fever over 104 degrees |
| 4. | patient | have | fever over 102 degrees |
| 5. | patient | have | sickle cell disease |
| 6. | patient | have | *another episode of *pneumonia* |
| 7. | patient | have | *brief period of *stiff* neck |
| 8. | patient | have | *prolonged *febrile* period |
| 9. | patient | have | *reduction of *fever* |
| 10. | patient | have | +temperature elevation to 101 |
| 11. | patient | have | +minimal neurological findings |
| 12. | patient | have | +progressive recovery |

*Instances of the computed-attribute construction, with the SIGN-SYMPTOM adjunct in italics.
+Instances of other selectional patterns.

Fig. 3: Object noun phrases of the verb *have* with the subject *patient*.

In this manner, we can also obtain a class of PATIENT words which occur as the subject of a V-PT verb, with a SIGN-SYMPTOM object (shown in Fig. 4). Examples of the PATIENT class from this corpus include:

*patient (patient developed fever);*
*she (she developed seizures);*
*he (he has sickle cell disease);*
*female (this 14 month black female, who is known to have sickle cell disease);*
*son (it was learned that 1 son had sickle cell disease);*
*sib (one sib, now dead in car accident, also had sickle cell disease);*

In some cases, e. g., *she developed painful hands*, the object noun phrase as a whole *(painful hands)* is in the SIGN-SYMPTOM class, although the head noun is not. This is an instance of a noun phrase whose class is determined by the class of one of its modifiers, rather than by the head noun. Because of the computational solution developed for this case, we call this the computed-attribute construction. In the above example, *painful* is in the SIGN-SYMPTOM class because of its morphological relation to *pain*; the combination of the left adjunct from the SIGN-SYMPTOM class *(painful)* with a BODY-PART head noun *(hands)* computes to a SIGN-SYMPTOM noun phrase (*painful hands, stiff neck, broken leg*, etc.).

Another case of the computed-attribute construction accounts for an apparent anomaly in the distributional data. We find such occurrences as

*patient had prompt reduction of fever* and *patient had brief period of stiff neck*, which we can relate to *patient have fever* and *patient have stiff neck* by recognizing that *reduction* and *period* are members of classes (respectively the CHANGE and TIME-PERIOD classes) that occur characteristically in the computed-attribute construction.

| | SUBJECT (HUMAN) | VERB | OBJECT |
|---|---|---|---|
| 1. | he | have | sickle cell disease |
| 2. | she | have | chicken pox |
| 3. | she | have | aplastic crisis |
| 4. | she | have | sickle cell disease |
| 5. | female | have | sickle cell disease |
| 6. | sibling | have | sickle cell disease |
| 7. | girl | have | *two episodes of *meningitis* |
| 8. | she | have | *two episodes of *meningitis* |
| 9. | female | have | *several *loose watery* stools |
| 10. | she | have | †no bowel movement for 2 days |
| 11. | parent | have | †trait |
| 12. | she | have | †measles vaccine |
| 13. | she | have | †primary immunization series |

*Instances of the computed-attribute construction, with the SIGN-SYMPTOM adjunct in italics.
†Instances of other selectional patterns.

Fig. 4: Human subjects (other than *patient*) of the verb *have*.

Figure 3 shows the distribution of the verb *have* with the subject *patient*; Fig. 4 shows the data for *have* with a human subject other than *patient*. The first nine occurrences in both Figs. 3 and 4 have a SIGN-SYMPTOM object or a computed-attribute SIGN-SYMPTOM object; the remaining occurrences are instances of less frequently occurring patterns.

We now examine a different environment which involves a number of new subclasses, in addition to the SIGN-SYMPTOM, PATIENT, and V-PT subclasses built up so far. The verb *show* has a large number of occurrences in the texts, shown in part in Fig. 5. The set of subjects can easily be grouped by inspection into two classes: a BODY-PART, class, consisting of parts of the body (*throat*, *abdomen*, etc.) or body substances (*spinal fluid*, *cerebrospinal fluid*); and a LAB class, consisting of laboratory tests (e. g., *urinalysis*, *x-ray*).

At first glance, the distribution of the objects of *show* appears to be fairly scattered among several different classes. On further examination an interesting pattern emerges. The classes LAB-RESULT and SIGN-SYMPTOM can be grouped together as the RESULT of an observation or a test made about the patient state. Depending on the source of the information



| | SUBJECT | VERB | OBJECT | SUBJECT CLASS | OBJECT CLASS |
|---|---|---|---|---|---|
| 1. | ( ) | show | severe anemia be due to sickle cell disease | ( ) | sentence |
| 2. | abdomen | show | no organomegaly | BODY-PART | SIGN-SYMPTOM |
| 3. | cerebrospinal fluid | show | WBC (be) 5022 | BODY-PART | sentence |
| 4. | cerebrospinal fluid | show | WBC (be) 55 | BODY-PART | sentence |
| 5. | chest | show | slight intercostal retractions | BODY-PART | SIGN-SYMPTOM |
| 6. | extremities | show | minor tenderness on extension of left wrist | BODY PART | SIGN-SYMPTOM |
| 7. | spinal fluid | show | no WBC | BODY PART | LAB-RESULT |
| 8. | throat | show | no pathogens | BODY-PART | LAB-RESULT |
| 9. | urinalysis' | show | no persistent abnormalities | LAB | SIGN-SYMPTOM |
| 10. | chest x-ray | show | density in left upper lobe | LAB | SIGN-SYMPTOM |
| 11. | chest x-ray | show | evidence of involvement of lingula as well as left upper lobe | LAB | ? |
| 12. | chest x-ray | show | infiltrate in RLL | LAB | SIGN-SYMPTOM |
| 13. | chest x-ray | show | densities in left upper lobe and lingula | LAB | SIGN-SYMPTOM |

Fig. 5: Subjects and objects of the verb *show* (with their word classes on the right).

(a LAB test, an EXAM-TEST made during a physical examination, etc.), we will get one of these types of RESULT. However this still leaves unexplained the alternation between BODY-PART and LAB in the subject of *show* and related verbs.

We find a clue to this alternation in the BODY-PART modifiers of some of the LAB words (e. g., *chest x-ray*); we also note that the word *urinalysis* has the implicit BODY-PART *urine*: in short, many of the LAB words have an associated BODY-PART word. On the other hand, the BODY-PART subjects seem often to have an implicit TEST word. This is particularly clear with pairs of sentences like 1 and 2 below, where *x-ray of spine* alternates with *spine* as the subject of *show*:

1. X-rays of spine show extreme arthritic change.
2. The dorsal spine shows moderately severe degenerative changes about the interspaces.

This example indicates that the alternation is a result of the application of a sublanguage-specific transformation; this transformation allows substitution of a modifier of a certain class (e. g. a BODY-PART modifier) for a head noun of another class (LAB here). In this example, the construction LAB OF BODY-PART is replaced by BODY-PART: *x-rays of spine* → *spine*. We can reconstruct the missing TEST word from the type of RESULT in the object, and from its immediate context within the document (e. g., the occurrence of the sentence in the radiology data, or in the physical examination section, etc.). If the object of *show* is a LAB-RESULT, the omitted test may be a culture *(throat shows no pathogen → throat [culture] shows no pathogen);* if the RESULT is a SIGN-SYMPTOM word, the missing TEST is the test appropriate to the part of the body done during the examination (e.-g., for *abdomen shows no organomegaly, palpation* is the appropriate test: *abdomen [palpation] shows no organomegaly*).

This pattern of distribution for the verb *show* is also seen for certain other verbs, e. g., *reveal (lungs reveal bilateral rhonchi), confirm (repeat cultures confirm a pathogen), be positive for (blood cultures are postitive for pneumococcus),* and *suggest (chest x-rays suggest pleural effusion).* On this basis, we group these verbs together into the V-SHOW class. The distributional data for these verbs is shown in Fig. 6.

Following the procedure sketched above, we obtain a set of word classes for the sublanguage. The distributional data characteristic of the sublanguage can then be expressed compactly as sequences of these word classes. For our corpus of hospital discharge summaries, we have obtained approximately 50 classes, although this number reflects certain classes where the distributional data was augmented with semantic refinements needed for retrieval of formatted information. The complete set of word classes for the clinical reporting sublanguage is shown in Appendix A.



| | SUBJECT | VERB | OBJECT | SUBJECT CLASS | OBJECT CLASS |
|---|---|---|---|---|---|
| 1. | lung | reveal | bilateral rhonchi | BODY-PART | SIGN-SYMPTOM |
| 2. | laboratory studies | reveal | anemia | LAB | SIGN-SYMPTOM |
| 3. | laboratory studies | reveal | systemic infection | LAB | SIGN-SYMPTOM |
| 4. | PES visit | reveal | *no source of infection | * | * |
| 5. | chest x-rays | reveal | progressive bilateral pneumonia | LAB | SIGN-SYMPTOM |
| 6. | study of pre-translusion specimen | reveal | *positive sickle cell preparation | LAB | * |
| 7. | ( ) | confirm | *impression of meningitis by finding cloudy CSF | ( ) | SIGN-SYMPTOM |
| 8. | repeat cultures | confirm | a pathogen | LAB | LAB-RESULT |
| 9. | examinations | confirm | bilateral pneumonia | EXAM-TEST | SIGN-SYMPTOM |
| 10. | chest x-ray | confirm | bilateral pneumonia | LAB | SIGN-SYMPTOM |
| 11. | blood cultures | be positive for | pneumococcus | LAB | LAB-RESULT |
| 12. | chest x-rays | suggest | pulmonary infarct | LAB | SIGN-SYMPTOM |
| 13. | chest x-rays | suggest | pleural effusion | LAB | SIGN-SYMPTOM |

*Instances of the computed-attribute construction
²Instances of other selectional patterns

Fig. 6: Occurrences of verbs similar to *show* with the sublanguage word classes of subject and object.

## 2.2. Designing the Information Format

An information format is a convenient representation of the major sublanguage word class sequences. To facilitate information retrieval applications, the word class sequences are formed into one composite structure, represented as a table with a column for each major word class. Within a row of the information format table, we recognize the basic sublanguage word class sequences by noting which columns are filled. A partial information format, based on the co-occurrence patterns discussed earlier, is shown in Fig. 7 with three formatted sentences. Sentences 1 and 2 *(Patient has sickle cell disease, She developed painful hands)* illustrate the PT V-PT SIGN-SYMPTOM pattern, while sentence 3 *(Chest x-ray revealed progressive bilateral pneumonia)* illustrates the BODY-PART TEST V-SHOW SIGN-SYMPTOM pattern.

As discussed in the previous section, a pattern larger than the subject-verb-object pattern is needed to capture the information patterns of the sublanguage. The examples in Fig. 5 showed the alternation between TEST and BODY-PART in the subject position of V-SHOW *(chest shows retractions* vs. *chest x-ray shows density)*. We accounted for this alternation by building a larger pattern, namely:

BODY-PART TEST V-SHOW RESULT

Those sentences with a TEST in the subject position generally either have a BODY-PART modifier *(chest x-ray)* or have a BODY-PART implicit in the TEST word (*urinalysis* implies BODY-PART = *urine*). In the sentences with BODY-PART subject, the TEST may be present in an adjunct, e. g., *lungs clear to auscultation*, or reconstructible from context. Figure 8 shows five sentences that differ greatly in syntax but all share the information pattern BODY-PART TEST [V-SHOW] RESULT.

These sentences illustrate the distinction between English paraphrastic relations and sublanguage paraphrase. These five sentences cannot be reduced to the syntactic sequence

TEST of BODY-PART V-SHOW RESULT

by any standard set of English transformations, but they are reduced to parallel representations in the format, based on the sublanguage class membership of the sentence words. For example, we can paraphrase *no abdominal masses felt* by the sentence *feeling [i. e., palpation] of the abdomen showed no masses*. However this relationship is not an instance of a sequence of general English transformations. It might be possible to define this sublanguage paraphrastic relation in terms of a series of sublanguage-specific transformations, including, for example, a transformation that allowed movement of a BODY-PART prepositional modifier from a TEST noun to a SIGN-SYMPTOM noun:

Palpation of the abdomen showed no masses →
Palpation showed no masses of the abdomen



| | PATIENT | PT-STATUS | | FINDING | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | V-PT | | BODY-PART | TEST | | V-SHOW | QUAL | | |
| | | | | | LAB | EXAM-TEST | | SIGN-SYMPTOM | LAB-RES | |
| 1. | patient | has | | | | • | | sickle cell disease | | |
| 2. | she | developed | | hands | | | | painful | | |
| 3. | | | | chest | x-ray | | revealed | progressive bilateral pneumonia | | |

Sentence 1: Patient has sickle cell disease.
Sentence 2: She developed painful hands.
Sentence 3: Chest x-ray revealed progressive bilateral pneumonia.

Fig. 7. Partial information format with three formatted sentences.

| | PATIENT | PT-STATUS | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | V-PT | BODY-PART | FINDING | | | | | | | | | | |
| | | | | | TEST | | V-SHOW | NEG | RESULT | | | | | |
| | | | | | LAB | EXAM-TEST | | | NORMALCY | QUANT | QUAL | | | |
| | | | | | | | | | | | SIGN-SYMPTOM | LAB-RES | | |
| 1. | | | [urine] | urin-alysis | | showed | no | | | abnormalities | | | | |
| 2. | | | lungs | | | revealed | | | | bilateral rhonchi | | | | |
| 3. | | | abdominal | | felt | | no | | | masses | | | | |
| 4. | | | liver | | palpable | | | | | 4 cm | | | | |
| 5. | | | right lung | | to per-cussion | | | | clear | | | | | |

[ ] Material in square brackets reconstructed from other entries in the row.

Sentence 1: Urinalysis showed no abnormalities.
Sentence 2: Lungs revealed bilateral rhonchi.
Sentence 3: No abdominal masses felt.
Sentence 4: Liver palpable 4 cm.
Sentence 5: Right lung clear to percussion.

Fig. 8: Partial information format illustrating syntactic variations in the TEST-RESULT relation.



However this transformation must be formulated in terms of sublanguage word classes. In addition, it can be considered an information-preserving transformation only in the context of this particular sublanguage.

We have chosen not to formulate these sublanguage relations into a set of sublanguage-specific transformations. This is because, at least for this sublanguage, it has been possible to map sentences into their format representation using a much simpler mechanism based on sublanguage class membership of the individual words. This is possible because within the sublanguage there is generally only one possible arrangement of a given set of word classes. For example, given the word classes SIGN-SYMPTOM, V-SHOW, and LAB, there is only one meaningful arrangement of these within the sublanguage, namely that the LAB test shows (V-SHOW) a SIGN-SYMPTOM. There is no other informationally distinct combination of these three classes in this sublanguage. Recognizing the unique combinatorial relations between subclasses allows us to by-pass a sublanguage transformational stage, and to use to a simpler mechanism for mapping into the format. This issue will be discussed further in the section describing the information formatting procedure (section 3.3).

Note that in Fig. 8 the various sources of information (e. g., LAB, EXAM-TEST) are grouped together under the heading TEST. Similarly the various results (NORMALCY, QUANT(itative), QUAL(itative)) are grouped under the heading RESULT. The heading QUAL in turn groups together several types of qualitative results: SIGN-SYMPTOM, and LAB-RESULT. These headings simplify retrieval and also facilitate the statement of certain larger patterns (e. g. TEST V-SHOW RESULT).

## Expanding the Format

Next, let us look at a pattern that we cannot accomodate in the partial format developed so far. Figure 9 shows the occurrences of *give* and some related verbs from the V-TREAT class. The basic pattern for *give* is

| SUBJECT | VERB | INDIRECT-OBJECT | DIRECT-OBJECT |
|---|---|---|---|
| ( ) | give | PATIENT | RX |

where RX is either a drug, some kind of treatment (e. g. *mist, transfusion*), or the classifier noun *treatment* or *therapy*. The verb *provide* has a distribution similar to *give*. The word *treat* shows a related pattern, with DIRECT OBJECT = PATIENT and PREPOSITIONAL OBJECT = RX: *( ) treat patient with drug. Treat,* however, has an interesting occurrence which does not fit into this pattern. It occurs with a disease (SIGN-SYMPTOM) object *(treat tonsillitis)*, as well as with a PATIENT object. This indicates that *treat* acually has three objects:

DOCTOR treat PATIENT for SIGN-SYMPTOM with RX

| | DOCTOR SUBJECT | VERB | PATIENT OBJECT | SIGN-SYMPTOM OBJECT | RX OBJECT |
|---|---|---|---|---|---|
| 1. | ( ) | give | her | | folic acid |
| 2. | ( ) | give | her | | phenobarbital |
| 3. | ( ) | give | her | | valium |
| 4. | ( ) | give | patient | | penicillin injection |
| 5. | ( ) | give | ( ) | | oxygen therapy |
| 6. | ( ) | give | ( ) | | mist |
| 7. | ( ) | give | ( ) | | transfusion |
| 8. | ( ) | give | ( ) | | asymptomatic treatment |
| 9. | ( ) | give | ( ) | | treatment of ampicillin |
| 10. | ( ) | provide | ( ) | | bed rest |
| 11. | ( ) | provide | ( ) | | hydration |
| 12. | ( ) | provide | ( ) | | partial exchange transfusion |
| 13. | ( ) | treat (with) | patient | | ampicillin |
| 14. | ( ) | treat (with) | | tonsillitis | ampicillin |
| 15. | ( ) | control | | seizures | valium |
| 16. | ( ) | control | | seizures | phenobarbital |

Fig. 9: Occurrences of the verbs *give*, *provide*, *treat*, and *control*.

This pattern enables us to build a still larger pattern. We note that the pieces PATIENT and SIGN-SYMPTOM have already been encountered in another pattern, namely PATIENT V-PT SIGN-SYMPTOM. Fitting these two pieces together, we build the larger pattern

| MD | TREATMENT | | PATIENT | PATIENT-STATE | | | | |
|---|---|---|---|---|---|---|---|---|
| | V-TREAT | RX | | V-PT | BODY-PART | TEST | V-SHOW | RESULT |
| | | | | | | | | |

Note that the structure of the information format allows us to collapse several distinct patterns, such as the PATIENT V-PT SIGN-SYMPTOM and MD V-TREAT PATIENT RX SIGN-SYMPTOM patterns. It also allows us to reduce to a single format representation the varying syntax of the objects of the V-TREAT verbs. For example, *give* takes PATIENT as the indirect object, and RX as the direct object *(give a drug to a patient)*. *Treat* on the other hand can take a PATIENT direct object, and RX as a prepositional object *(treat a patient with a drug)*. However, in the format, this syntactic variation is replaced by an alignment of words depending on their sublanguage word class membership, as illustrated in the partial format of Fig. 10.



| | MD | TREATMENT | | PATIENT | PT-STATUS | | FINDING | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | V-TREAT | RX | | V-PT | BODY-PART | TEST | | | V-SHOW | NEG | RESULT | | | | |
| | | | | | | | LAB | EXAM-TEST | | | | NORMALCY | QUANT | QUAL | SIGN-SYMPTOM | LAB-RES |
| 1. | | give | penicillin injection | patient | | | | | | | | | | | | |
| 2. | | treat with | ampi-cillin | | | [tonsils] | | | | | | | | | tonsillitis | |
| 3. | | control | valium | | | | | | | | | | | | seizures | |

Sentence 1: Patient was given a penicillin injection.
Sentence 2: Tonsillitis was treated with ampicillin.
Sentence 3: Seizures were controlled by valium.
[ ] Material in square brackets reconstructed from other entries.

Fig. 10: Extending the information format to cover TREATMENT expressions.

## Time, Aspectual, and Modality Operators

To complete the construction of the information format, we add columns for the representation of certain general classes of operators such as time and aspectual operators, as well as negation and uncertainty markers. These operators are distinguished by the fact that they are general to English and are *not* specific to this sublanguage. Therefore they show different distributional patterns, specifically a wider distribution than the sublanguage-specific classes. For example, BODY-PART adjectives in this corpus modify only three classes of words in the sample eight-document corpus. By contrast, negation modifies eleven classes. Similarly, the V-SHOW class occurs with four subclasses in subject position; the CHANGE class on the other hand occurs with ten different subclasses of subject.

The information format is constructed to reflect the wide distribution of these operators. Whereas the sublanguage-specific classes in this material are each represented by a single format column, the general English classes have several identical sets of columns in the format, one set associated with each sublanguage verb or event. That is, there is a set of operator columns associated with V-PT, a set with V-SHOW, a set with V-TREAT, a set with RESULT, etc. In this way the format represents the relation between the operator and the word that it operates on, e. g., *2 day treatment with penicillin for fever* vs. *was treated with penicillin for fever of 2 days*. Figure 11 shows a partial information format with the representations of these two sentences. The time modifier in the first sentence is associated with V-TREAT, under the new heading VERB-TREAT; the time modifier in the second sentence is associated with RESULT.

Despite the wide distribution of English operators, not all sublanguage classes take negation or time modifiers. For example, the PATIENT and INST(itution) classes and the BODY-PART class do not. This reflects certain specific informational characteristics of this particular sublanguage: PATIENT and INST do not take these modifiers because a hospital discharge summary contains information about a particular patient during a particular hospitalization at a given hospital; it will not contain information about who is not a patient or what is not a hospital. The BODY-PART class is not generally negated or modified because the BODY-PARTs are expected to be present. Thus the sublanguage does not contain sentences such as *patient had a neck for two days*, although we can have *patient had a stiff neck for two days*, where the time modifier modifies the computed-attribute SIGN-SYMPTOM phrase *stiff neck*.

In general, those subclasses that can take one kind of operator in this set can take others as well. Therefore a set of columns for these operators is provided at each of seven places in the format. The columns for the operators are grouped under two headings. The first, MODS, has columns for negation (NEG), for modal operators of possibility or uncertainty



| | MOD | TREATMENT | | | | PT | PT-STATUS | | BODY-PART | FINDING | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | VERB-TREAT* | | | | | VERB-PT* | | | TEST | VERB-SHOW* | | RESULT* | | | | |
| | | V-TREAT | TIME* | | RX | | V-PT | TIME* | | | V-SHOW | TIME* | NORMALCY | QUANT | QUAL* | TIME* | EVENT-TIME |
| 1. | | treatment with | 2 day | EVENT-TIME | penicillin | | | | | | | | | | | fever | |
| 2. | | treatment with | | | penicillin | | | | | | | | | | | fever | of 2 days |

Sentence 1: 2 day treatment with penicillin for fever.
Sentence 2: Treatment with penicillin for fever of 2 days.
*Only the EVENT-TIME column is shown under TIME to conserve space.
'Verb or event columns that take TIME (and MODS, not shown here).
•Sub-headings are not shown for these columns.

Fig. 11: Partial information format illustrating placement of TIME operators within the format.

(MODAL), e. g. *suggest* as in *x-ray suggested pulmonary infarct* or *possible* in *possible infection*. The MODS heading also includes a column (EVID) for assertional operators: operators which assert the existence of their operand e. g., *evidence* in *x-ray showed evidence of metastasis*, or *present* in *nasal discharge present*. This is summarized in the top half of Fig. 12.

| MODS | | |
|---|---|---|
| NEG | MODAL | EVID |
| not<br>fail to | may<br>possible | evidence of<br>occur |

typical MODS entries

| TIME | | | | | | |
|---|---|---|---|---|---|---|
| EVENT-TIME | V-TENSE | CHANGE | ASP | | | REP |
| | | | BEG | END | TM-PER | |
| 1 day ago<br>now<br>at discharge | past<br>perfect<br>future | increase<br>fluctuate<br>unstable | appear<br>new<br>start | complete<br>through<br>finish | continue<br>period<br>still | repeat<br>twice<br>often |

typical TIME entries

Fig. 12: MODS and TIME columns, showing their substructure and sample entries.

The second group of columns are collected under the heading TIME. This heading includes a set of columns (under the heading EVENT-TIME) for expressing time point expressions or durational expressions (e. g., *on the day of admission, for two days*). TIME also contains a column (V-TENSE) for verb tense, a CHANGE column (for CHANGE words such as *increase, rise* etc.); it contains a heading ASP(ectual) with the columns BEG(in), END, and TM-PER (for continuation over time, e. g. *continue* or *remain*). Finally there is the column REP under TIME, for expressions of repeated occurrence (e. g. cardinal numbers greater than one, and words such as *frequent, occasional, repeated*, etc.). The TIME heading and its subparts is illustrated in the bottom half of Fig. 12. For a more detailed account of these operators, see [Sager and Hirschman 1978].

The representation presented here does not deal with the issue of the relative *scope* of these operators. In this material it has been possible, in general, to ignore this problem because there is rarely more than one such operator in a sentence. In this case it is sufficient to record the word it



operates on. However in the case where there would be more than one operator, this formalism is not adequate to represent the possible differences in information, as in *possible change in a lesion* vs. *change in a possible lesion*.

## 3. *Sublanguage Document Processing*

### 3.1. Parsing

#### Functions of the Parse

For a sublanguage which involves any degree of linguistic complexity, obtaining a parse is the first step in determining the informational content of the sentence. The parser, using an appropriate grammar and lexicon, computes from the linear ordering of words in the input string a set of syntactic relations that hold between the words of the sentence and enable the string of words to be recognized as a grammatical sentence of the language.

Possibly there exist sublanguages that are so simple that utterances in the sublanguage are still intelligible no matter what the order of the words in the utterance. This kind of situation can occur if there are no homographs in the sublanguage (so that each word has a unique informational class), and further if there are no syntactically ambiguous constructions, e. g. no conjunctions, or nested modifiers. Such sublanguages would not require the linguistic processing described here because of their highly restricted nature.

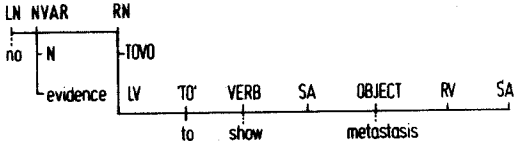
In the case where the input is continuous text, even a short text composed of short sentences, there is reason to provide a syntactic analysis as a step toward the informational representation of the sentence. (See Appendix B for a text sample from the sublanguage of clinical reporting). Some of the syntactic information available from the parse is needed to assign correct modifier-host relations to common English modifiers that carry important information, such as negation and time expressions. Thus, *no pulse* is very different from *no pulse taken*; and *Had 1 day fever prior to admission* has different time information from *Had fever 1 day prior to admission*. One also needs the parse to establish very basic syntactic relations, such as the subject and object of the verb, which carry important information (cf. *Mother reported patient had sickle cell disease* vs. *Patient reported mother had sickle cell disease*).

We also need a parse because of ever-present syntactic ambiguity. The parse sometimes resolves a potential ambiguity; more often it simply lays forth the possible readings so that sublanguage constraints can be brought to bear, either to eliminate unlikely readings, or to select the most likely one. This can be illustrated by a simple example of a typical ambiguity due to conjunction. In a sequence such as *Patient received two injections and transfusion*, the grammatical constraint that a quantifier and its host noun agree in number, applied during parsing, rules out the wrong

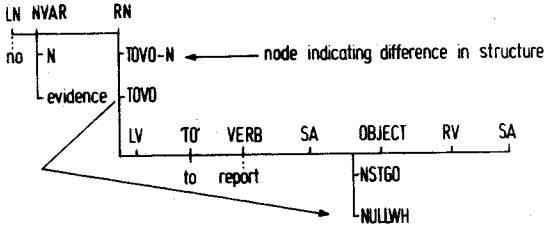
reading (the one which would have the patient receive two transfusions). However, in *Patient received two injections of morphine and transfusion*, grammatical rules alone will not eliminate the reading corresponding to *Patient received two injections of transfusion*. For this we would need to know that *transfusion* is not the proper type of sublanguage modifier for the noun *injection*. What the parse can do is to set up the structural possibilities (e. g. alternative parses for the sublanguage word string) in such a way that sublanguage selectional constraints can be applied to the appropriate sequences.

The parse also sets up the structures in relation to which certain implicit, or "zeroed", word occurrences can be restored. Almost every conjunctional construction (except those consisting of conjoined full assertions) is of this type. Thus in the above example, from *patient received two injections and transfusion* we obtain *patient received two injections and patient received transfusion*. The grammatical rules for the expansion around conjunctions are embodied in a conjunction expansion transforma-

Sentence 1: No evidence to show metastasis.




Sentence 2: No evidence to report.




The arrow points to the missing OBJECT of TOVO-N (to be copied from host = *evidence*)

Definition of symbols:

| | |
|---|---|
| LN | = left noun adjunct; |
| LV | = left verb adjunct; |
| NVAR | = noun or variant |
| RN | = right noun adjunct; |
| RV | = right verb adjunct; |
| SA | = sentence adjunct; |
| TOVO | = *to* + V(erb) + O(bject); |
| TOVO-N | = *to* + V(erb) + O(bject) − N(oun phrase); |
| NSTGO | = N(oun) ST(rin)G O(bject); |
| NULLWH | = "missing" noun phrase |

Fig. 13: Parse trees for two noun phrases with slightly different analyses.



tion which is applied to the parse tree of the sentence. Thus another function of parsing is to provide the relevant structures for the operation of transformations, discussed in greater detail in section 3.2.

An additional example serves to illustrate how the detailed surface structure description provided by the parse facilitates later transformational regularization. The noun phrase *no evidence to show metastasis* resembles superficially the noun phrase *no evidence to report*. They differ, however, in that the head noun *evidence* is the implicit subject in one case *(no evidence to show metastasis = no evidence such that evidence shows metastasis);* in the other case, it is the implicit object: *no evidence to report = no evidence such that ( ) report evidence*[2]. The parse must produce these different analyses (shown in Fig. 13) in order to apply subject-verb-object selectional restrictions correctly to these two structures. The first has the right noun adjunct (RN) of the type TOVO (TO + Verb + Object). The second has the RN of type TOVO-N (TO + Verb + Object − Noun phrase), indicating that the head noun is the "missing" object noun phrase of the TOVO string. The differing parses allow a very simple statement of the transformations required to convert the right adjuncts into full ASSERTIONs, since the transformations do not have to re-analyse the structure to determine what type of regularization to apply.

## Parsing Problems in the Sublanguage of Clinical Reporting

Documents in the sublanguage of clinical reporting are distinguished syntactically in various respects. Most striking is that the document "sentences" are often not complete sentences at all. The sentential unit, i. e. the stretch between end-of-sentence punctuation marks, may indeed be a full assertion *(The gram stain was negtive);* it may also consist solely of a noun phrase *(No contact with known infection)*, a subject + predicate *(conjunctivae clear)*, an assertion lacking only a subject *(will be followed in clinic)*, or certain other forms as illustrated in Fig. 14. It may also consist of a sequence of such forms separated by commas or other punctuation. These forms can be related to the full assertion form by positing that they are derived by the deletion of grammatical function words (chiefly *be*) which appear as droppable constants in many transformations, or by the deletion of words which have distinguished status within the sublanguage. An example of the latter in the clinical sublanguage is the absence of the word *patient* in the subject position where the verb is a V-PT type: *Had previous hospitalization*. Another example is the absence of particular TEST words when the noun which is present implies the test, e. g. in a physical-examination paragraph *No liver* in place of *No liver felt*.

[2] Also the modalities of the verbs differ in these two constructions, the second verb having a future or non-actual time associated with it.

| | |
|---|---|
| GERUND: | *finally resolving.* |
| INFINITIVE: | *to be followed in medical clinic;*<br>*to be followed in clinic by Dr. Lieberman;*<br>*to continue work-up of anemia.* |
| SUBJECT-LESS ASSERTION: | *was seen by local MD and was treated with erythromycin, which she took for 5 days.* |
| SUBJECT-PREDICATE: | *conjunctivae clear;*<br>*platelets adequate;*<br>*pupils equal and reacting to light;*<br>*reflexes present rated 2+;*<br>*temp. 100; pulse 90 and reg; res. 24; BP 110/60.* |
| NOUN PHRASE: | *tonsillectomy as a child;*<br>*no hypertension, diabetes, heart or kidney disease;*<br>*no contact with known infection.* |
| ADJECTIVE PHRASE: | *dry without exudate.* |
| PREPOSITIONAL PHRASE: | *in no distress.* |
| PASSIVE STRING: | *loaded with gram positive cocci in pairs.* |

Fig. 14: Examples of sentence fragments in clinical reports.

The automatic parsing of medical documents is being done using the NYU Linguistic String Project system. Its large grammar of English [Sager 1981] has been adapted for the medical sublanguage by adding to the set of constructions defined in the general grammar of English, the definitions of "fragments" (the elliptical forms shown in Fig. 14), and by modifying the restriction component (constraints on the parse tree) to account for grammatical deviations from standard English, e. g., the frequent absence of articles [Anderson et al. 1975].

Restrictions have also been added to disallow unwanted parses. As is well known in automatic parsing, the machine often "discovers" readings of the sentence which are permitted by the grammar but are not the intended reading. An example involving the assignment of an adjunct to the correct noun of two conjoined nouns was given earlier *(injections of morphine and transfusion)*. The introduction of freely combining sentence fragments into the grammar compounds the ambiguity problem. Add to this the pervasive inconsistency in the use of punctuation in the texts and the use of the comma for multiple functions, and the result is a parsing problem of considerable magnitude.

Solutions to the medical document-parsing problem have proceeded along several lines. First, we edit the grammar and the dictionary, which were both designed for English texts and therefore include structures and word usages that are not likely to be encountered in patient records. This is illustrated in Fig. 15, parts A and B (from [Insolio and Sager 1977]) which



A. Edit the too-rich English dictionary for this sublanguage:
   1. *Eating well.* (cf. *wishing well*)
      Remove the noun class from *well.*
   2. *She remained well throughout 14 days of penicillin therapy.*
      *She responded well to penicillin for 2 weeks.*
      (cf. *The meeting continued well into the night.*)
      Remove left-of-preposition subclass from adverbial class of *well.*
   3. *Patient shows change since last week.*
      (cf. *Movie shows change from week to week.*)
      Remove the NOUN class from *show(s).*

B. Edit the too-rich English grammar for this sublanguage:
   WRONG PARSE: *Penicillin when there is evidence of infection.*
   (cf. Penicillin, when there, is evidence of infection.)
   Remove string covering *when* + adverb from grammar, or require commas.

C. Selectional restrictions needed:
   1. *Seen in PES and admission advised.* [PES = Pediatric Emergency Service]
      (cf. *Seen in sentences and fragments parsed.*)
      Disallow conjoining of *PES* and *admission* in the sublanguage.
   2. *Heart grade 2/6 systolic murmur along left sternal border.*
      (cf. *Shoe size 8½ nice fit beneath metatarsal arch.*)
      Constrain compound noun formation to disallow e. g., *heart grade.*

Fig. 15: Sublanguage adjustment of dictionary and grammar.

shows examples that would receive the wrong parse if the suggested editing were not done. Second, we try to order the options of the grammar so as not to accept fragments if a fuller form is constructable from the same components. We also restrict the acceptance of fragment forms both with regard to syntactic composition and sublanguage word class. Lastly (to the extent possible during parsing), we apply selectional constraints to the forms most productive of spurious readings, chiefly conjunctional constructions. We define equivalence classes of sublanguage word classes and allow conjoining of words from the sublanguage vocabulary only if they are in the same equivalence class. In this regard, the notion of "computed attribute" mentioned in section 2.1 is important. A phrase such as *stiff neck*, whose head noun is in the BODY-PART class, must be assigned the SIGN-SYMPTOM attribute carried by the adjective so that it can conjoin with the SIGN-SYMPTOM nouns, as in *stiff neck and fever.*

To illustrate the output of the parsing, a portion of a discharge summary is shown in Appendix B and the parse tree obtained for one of the sentences in Appendix C.

### 3.2. English Transformational Decomposition

The function of the English transformations is to regularize the structure of the input sentence, while preserving its informational content. The regularization performed by the transformations is general to English and not

specific to a particular sublanguage. It consists of certain rearrangements into standard subject-verb-object form, normalization of certain morphological variants such as tense and plural markers, and the filling in of zeroed material in subordinate clauses, sentence nominalizations, and conjoinings. The original syntax of an input sentence is thus "decomposed" into simpler, more regular and more complete units that are nonetheless informationally equivalent to the original sentence.

The transformations operate on the trees produced by the parsing phase, which represent the surface syntax of the input sentences. The output of the transformations is, in turn, also trees belonging to this set. As a result, the output of one transformation may serve as the input to another transformation.

The transformational component makes use of the same Restriction Language used to formulate grammatical restrictions [Sager and Grishman 1975], augmented by a set of tree-building operators (e. g. insert, delete, replace), and a sequencing mechanism to control the order of execution of the transformations [Hobbs and Grishman 1976].

This section will outline the set of English transformations that have been applied to the sublanguage of clinical reporting. It will also discuss certain classes of transformations that were *not* included, and the reasons for their exclusion. We can distinguish six groups of transformations:

1) Fragment regularization, which converts the parse of a FRAGMENT[3] into an ASSERTION[3] structure with certain missing pieces;
2) Morphological regularization, which includes converting tensed verbs into the infinitive plus tense marker, plural nouns into singular nouns plus plural marker, numbers into their digit representation, etc.;
3) Conjunction expansion, which fills out conjoined structures, for example, making two complete ASSERTIONs from a conjoined noun phrase in the subject or object;
4) Subordinate clause regularization, which includes filling in the antecedent in a relative clause, and filling in material omitted in certain truncated forms of subordinate clauses and modifiers;
5) Rearrangements of sentence pieces, such as passive to active and replacement of the dummy sentence subjects *it* and *there*;
6) Regularization of nominalized sentences by converting them into full (embedded) ASSERTIONs.

[3] These words are capitalized because they refer to the strings by these names, as defined in the parsing grammar. For a list of the FRAGMENT types, see Fig. 14. The ASSERTION consists of the string

SA SUBJECT SA TENSE SA VERB SA OBJECT RV SA

where SA stands for sentential adjunct, RV for right verb adjunct. See Appendix C for an example of a parse tree.



### Fragment Regularization

This set of transformations converts FRAGMENT strings into ASSERTION strings with empty SUBJECT, VERB, or OBJECT slots (depending on the type of FRAGMENT). For example, the subject-less assertion (e. g. *was seen by local MD* from Fig. 14) is converted into an ASSERTION string with an empty SUBJECT. The adjectival and prepositional phrase FRAGMENTs are converted into ASSERTIONs with empty SUBJECT and VERB = *be*. The subject-predicate construction is converted into an ASSERTION with an empty VERB if the OBJECT is a noun phrase, as in *chest x-ray – no evidence of metastasis*; otherwise the verb *be* is filled in: *conjunctivae clear → conjunctivae be clear*. Noun string FRAGMENTs are converted into the SUBJECT of an ASSERTION, with VERB and OBJECT empty.

### Morphological Regularization

In order to reduce the number of alternating morphological forms, words are replaced by their "canonical" form plus a marker to preserve the morphological information where necessary (e.g., tense or number marker). Plural nouns are thus replaced by the singular form plus a plural marker: *x-rays → x-ray plural*. In addition, spelling variants and abbreviations are replaced by the full form: *xray → x-ray*; CSF → *cerebrospinal fluid*. Numbers are replaced by their digit representation: *eleven → 11*; *third → 3rd*.

The regularization of tensed verbs is slightly more complex, since it involves replacement of several words (the tense-bearing auxiliary and the main verb), in addition to rearrangement of any modifiers positioned between these two verb forms: *she has not had fever → she not have [present perfect] fever*. This set of transformations includes an imperative transformation (transforming the imperative into an ASSERTION with an imperative modality marker); it includes the future *to* transformation *(she is to be seen in hematology → she [future] be seen in hematology)*; and it includes transformations for converting modals and auxiliaries (e.g. *will*, *may*, *do*) into modality or tense markers.

There are also several transformations to remove prefixes, converting them into separate morphemes. For example, *x-ray is unchanged → x-ray is not changed*. Certain adjectives and adverbs with a time prefix (e.g., *pre*) are converted into preposition plus morphologically related noun: *post-operative → post operation*. The word *without* also has a special transformation that converts it into *with* plus *no*: *without complications → with no complications*.

### Conjunction Expansion

Conjunctions are expanded into independent ASSERTIONs (which in turn are represented as separate entries or rows in the format table). This process

removes conjunctions from within the individual ASSERTION: given the sentence *patient developed cough and fever yesterday*, conjunction expansion gives *patient developed cough yesterday and patient developed fever yesterday*. Similarly for a sentence with a conjoined verb phrase: *Exudative tonsillitis was found and treated with ampicillin → Exudative tonsillitis was found and exudative tonsillitis was treated with ampicillin.*

The conjunction *or* under negation requires an additional adjustment. Given the sentence *there was no cough or respiratory distress*, the *no* is distributed over both noun phrases, and in the expansion, the *or* becomes *and* , to produce the expanded sentence: *There was no cough and there was no respiratory distress.* However, not all conjunctions are expanded to complete assertions. In particular, numerical range expressions are not: *Hematocrit varied between 27 and 34* is not expanded, nor is *She smoked 1 or 1 1/2 packs of cigarettes a day*. These are represented as ranges within the QUANT(ity) column of the format.

### Subordinate Clause and Adjunct Regularization

This set of transformations regularizes subordinate clauses and certain other noun modifiers, converting them into full ASSERTIONs. The relative clause transformation copies the head noun modified by the relative clause into its appropriate place in the subordinate ASSERTION, e. g., *the transfusion that the patient received → the transfusion such that the patient received transfusion*. The copy of the head noun added to the relative clause is marked, to preserve the information that it is a copy and not a new occurrence of the same noun.

Certain right modifiers of the noun behave in similar fashion: *The patient had several seizures requiring valium → The patient had several seizures such that seizures require valium*. The right modifier transformation checks for a verbal form, an adjective, an appositive construction, or a CONNECTIVE preposition (e. g. *due to, like*). If one of these constructions is found, the transformation copies the head noun into the subordinate clause as its SUBJECT, and either regularizes the verbal form or fills in the verb *be* in order to create a complete ASSERTION: *the father, a construction worker, → the father such that father be a construction worker.*

Certain subordinate clauses introduced by a subordinate conjunction can also be in the form of incomplete assertions; these are handled in the same way as the expansion of right modifiers, namely by copying the subject and filling in the verb *be* if necessary: *She had an episode of dyspnea while in Maryland → She had an episode of dyspnea while she be in Maryland* (see Appendix D). Note that the subject-copying for the subordinate clause regularization must apply to the surface subject. That is, this transformation must take place before certain rearrangement transformations



(specifically the passive) have been applied. If the passive is applied first, the wrong noun phrase may be copied:

She was examined by a local doctor while in Maryland.

is transformed by the passive transformation into:

A local doctor examined her while in Maryland.

which in turn would be incorrectly transformed by the subordinate clause transformation into:

A local doctor examined her while doctor be in Maryland.

Finally there is a transformation that transforms sentential adjuncts containing a verb into a subordinate clause modifying the object or subject noun phrase: *She was discharged improved → She (such that she be improved) was discharged.*

### Rearrangement Transformations

These transformations rearrange various pieces of the ASSERTION. The passive transformation converts a passive construction into an active one: *Patient was seen by a local MD → A local MD saw patient.* The *it* and *there* replacement transformations replace the dummy *it* or *there* in the SUBJECT by the head noun in the OBJECT position, with appropriate adjustment in the OBJECT: *It was decided to continue treatment → To continue treatment was decided; There was cardiomegaly noted → Cardiomegaly was noted.* These transformations create structures on which other transformations can operate; for example the passive can operate on the output of the *it* or *there* replacement transformations in the two examples above. In fact, the *it* and *there* transformations *must* operate before the application of the passive, the subordinate clause expansion, or the sentential adjunct expansion.

### Sentence Nominalizations

There is a set of transformations that convert sentence nominalizations into embedded ASSERTIONs. For example, the object of the verb *prevent* is a sentential object in the sentence *This prevented the patient from being discharged.* This sentential object is converted transformationally into a full ASSERTION *(patient be discharged)* in the OBJECT position: *this prevent [past] (patient be discharged).*

Another type of nominalization consists of a noun phrase with a nominalized verb as its head noun (e. g. *admission* from *admit*). The left and right adjuncts of the nominalized verb are generally the subject and/or object(s) of the underlying verb: *admission of the patient to the hospital → ( ) admit patient to the hospital.* This type of nominalization has not been

reduced to its underlying ASSERTION form in this material. There are several reasons for this. One is that the syntax of the nominalized-verb noun phrase resembles the syntax of a regular noun phrase. Compare *radiation treatment* with *radiation therapy*, where *treatment* is a nominalized verb and *therapy* is not. Clearly these two phrases are closely related in meaning as well as syntax. To perform a denominalizing transformation on *treatment* would destroy this parallelism; no real regularization of syntax is achieved by transforming noun phrases with nominalized verbs as head into their underlying ASSERTIONs.

The problem of transforming nominalized-verb noun phrases raises the issue of how "deep" to transform: how simple (or complex) should the structures in the transformational output be? Should all adjuncts be converted into full subordinate ASSERTIONs (e. g. *painful crisis* → *crisis such that crisis be painful*)? Should all morphological decompositions be applied (for example, *crisis be painful* → *crisis V? pain*, where the missing verb (V?) is a function of the head noun *crisis* and its modifier *pain*, perhaps *cause* in this case).

The answer to these questions is that the depth of transformation depends on the intended use of the output of the transformational decomposition. In the application under discussion here, the output of the transformational decomposition is mapped into the information format. The format has been constructed to preserve the co-occurrence patterns within a sublanguage sentence. Therefore to transform all adjuncts into separate ASSERTIONs defeats part of the intention behind the format, because it would cause a single sentence to be represented as a set of ASSERTIONs which would appear in the format as a sequence of sparsely filled format rows, rather than as a single more densely filled row. In the discussion of the occurrences of *show* and later of *treatment* in sections 2.1 and 2.2 respectively, we saw that the information structures were larger than the standard syntactic relations of subject-verb-object. Because of this fact, it is not desirable to transform all noun adjuncts into independent ASSERTIONs in this application. However, for other applications, the primary interest may be in obtaining a very limited number of highly regular syntactic patterns, rather than in preserving the larger informational patterns. In such cases, a deeper transformational analysis would be appropriate.

### 3.3 Mapping into the Information Format

The set of formatting transformations maps each parsed, transformed input sentence into the appropriate columns of the information format table. The formatting transformations represent the sublanguage-specific component in the processing. Both the parsing phase and the English transformational phase are general to English as a whole, with the exception of a limited number of sublanguage-specific rules (for example, the constraint



on conjoining compatible nouns, discussed in section 3.1). As a result, the formatting component serves two functions: it applies sublanguage selectional restrictions to determine appropriate modifier placement and to disambiguate sublanguage homographs (e.g., *cancerous growth* where *growth* is a SIGN-SYMPTOM vs. *bacterial growth* where *growth* is LAB-RES vs. *growth and development of the patient* where *growth* is a DEVEL(opmental) word). The second function of the formatting component is to map the sentence words into the appropriate format columns on the basis of their membership in the corresponding sublanguage word class. These two activities are actually carried out simultaneously.

The formatting component makes use of the same mechanism as the English transformational component. After each ASSERTION node, a copy of the information format (represented now as a FORMAT subtree) is attached as a sister node. The formatting transformations then move pieces of the ASSERTION tree into the FORMAT subtree. At the end of a normal formatting process for a given sentence, all the words should be found under the FORMAT subtree, with no words left under the original ASSERTION. If there are words left over under the ASSERTION, these are printed out in a special message, to indicate that the formatting of the sentence is incomplete (cf. the sentence in Appendix E).

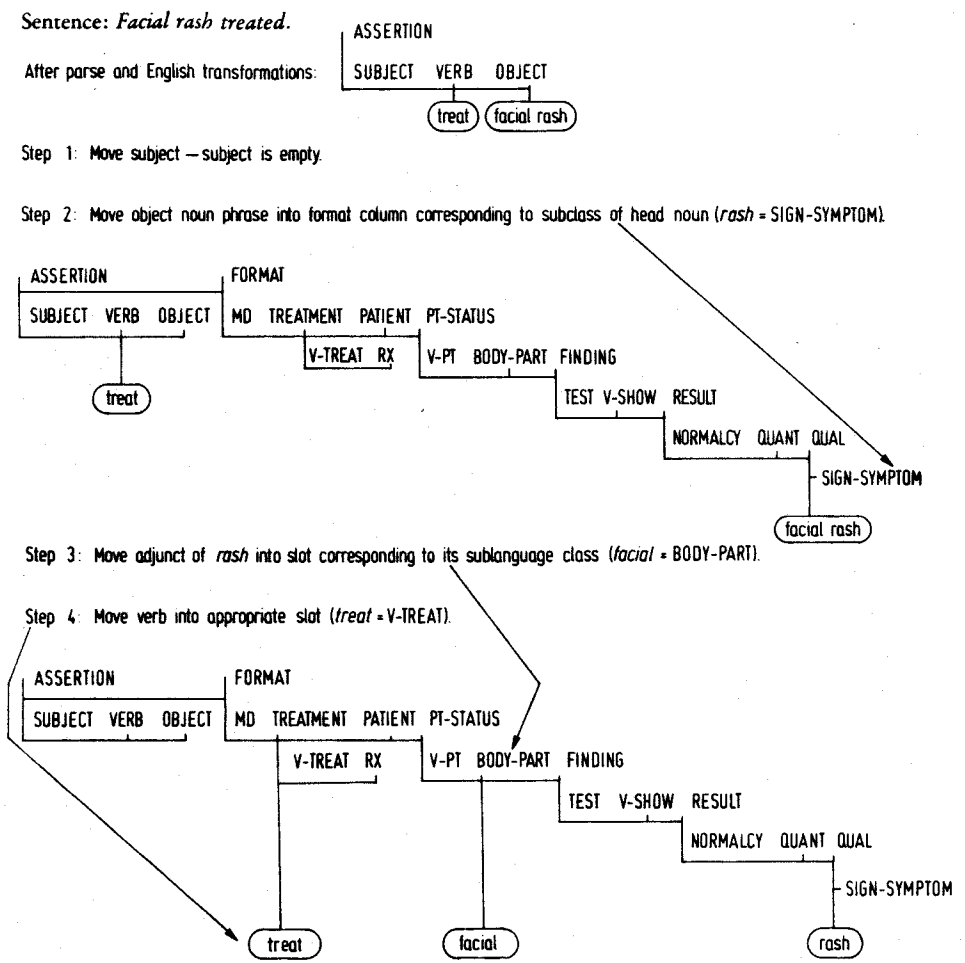
The words are moved into the FORMAT subtree in units of head plus modifiers; the sublanguage class of the head controls the mapping of the entire phrase into the corresponding format column (now represented as a node in the tree). Once the entire phrase has been moved into the FORMAT, the modifiers are then moved in similar fashion, if they have sublanguage classes corresponding to format columns. Otherwise they remain as left or right adjuncts on their original head noun, preserving their syntactic relation to it.

The elements of the ASSERTION are mapped into the FORMAT subtree in the order:

SUBJECT / OBJECT / VERB / SENTENCE ADJUNCTS

The verb is moved after the subject and object have been formatted, because its interpretation often depends on the informational role of the subject and object. For example, *be* can connect PATIENT and SIGN-SYMPTOM, like V-PT: *patient was sick*; or it can connect a TEST and a RESULT, like V-SHOW: *x-ray was normal*. Similarly it is easier to determine what the sentence adjuncts modify once the subject, verb, and object are all formatted. This process is illustrated for the sentence *facial rash treated* in Fig. 16.

Disambiguation of homographs is currently combined with the mapping of the words into the format. The transformations are organized so that there is a transformation for each subclass that corresponds to a format column. For example, there is a transformation for mapping a phrase with a




Note that the ASSERTION subtree now has no sentence words under it; they have all been moved to their appropriate places under the FORMAT subtree.
(TIME and MODS nodes omitted for brevity)

Fig. 16: Mapping a sentence into the format (shown here as a tree).

BODY-FUNC(tion) head noun into the BODY-FUNC column. In order to distinguish certain body functions from the same word used as a doctor's action (e. g. *patient hears well* vs. *doctor hears a heart murmur*), certain co-occurrence patterns are checked. Specifically, a BODY-FUNC verb cannot have a SIGN-SYMPTOM object. If it does, the BODY-FUNC transformation will not apply and a word like *hear* (in *hear heart murmur*) will be formatted by the EXAM-TEST transformation. The BODY-FUNC transformation also requires that the subject of the BODY-FUNC



verb (if present) be a PATIENT or a BODY-PART word (e. g., *moves* is a BODY-FUNC in *wrist moves normally*, but not in *doctor moved patient*). Note that these checks on the informational type of the subject and object are facilitated by the fact that when formatting the BODY-FUNC verb, the subject and object have already been mapped into the format. To check for a PATIENT subject, we check whether the subject has been moved into the PATIENT subtree[4]; the informational content of the subject or object has been determined by its location in the format.

There are several features of the formatting procedure that deserve explanation. One is the representation of sentences that contain more than one ASSERTION. This situation arises with any conjoining (since conjoined phrases are expanded into complete conjoined ASSERTIONs). It also arises for subordinate clauses, and in certain special cases, for "connective" operators, such as *cause*, or *be associated with*, which can connect two medical events, each requiring a separate row of the format table (since, in general, only one entry per column is allowed in each row).

In representing connected sentences, it is important to retain the information about how the sentences are connected; that is, to distinguish co-ordinate conjunction constructions with *and*, *or*, etc., from subordinate conjunction constructions with a particular conjunction such as *after* or *because*, from a relative clause or an embedded ASSERTION or a relation expressed by a connective word such as *cause*. It is also important to represent the relative scope of the connections correctly. For example, in a sentence such as *she has a fractured ankle and a sprained or fractured wrist*, the relative scope of the *and* and the *or* affect the meaning. The fully expanded sentence is ambiguous if no groupings are shown:

She has a fractured ankle
and she has a sprained wrist
or she has a fractured wrist.

To resolve these problems, a CONNECTIVE column is added preceding the FORMAT columns. This provides a place to enter information about the type of CONNECTIVE (CONJUNCTION / RELATIVE-CLAUSE / SUBORDINATE CLAUSE etc.) and the connector itself. By restricting the CONNECTIVEs to binary connectives and by placing the connector for two format rows in front of the first of the two formats that they connect (Polish notation), we obtain a unique representation for the connective and its scope. Thus the *fractured wrist* example would be represented as shown schematically in Fig. 17. The formatting transformations recognize the various types of connectives and map them into the

[4] Each formatting transformation sets links between the original position of a phrase in the ASSERTION subtree and its location in the format, so that it is possible to determine what columns the subject and object have been mapped into while formatting the verb.

CONNECTIVE column preceding the two FORMATs that they connect (see also Appendix E).

Sentence: *She has a fractured ankle and a sprained or fractured wrist.*

| CONNECTIVE | FORMAT |
|---|---|
| CONJUNCTION = and | she has a fractured ankle |
| CONJUNCTION = or | she has a sprained wrist |
| | she has a fractured wrist |

Equivalent logical notation:
(she has a fractured ankle) AND ((she has a sprained wrist) OR (she has a fractured wrist))

Fig. 17: Representation of connective scope in the information format

The relative clause transformation (and the related adjunct expansion transformation) present another problem. In regularizing the relative clause or in expanding an adjunct, the head noun has been copied into the subordinate clause. It is important to record that this is a copy of an element in the main clause and that it does not represent a separate occurrence of this event. In the current implementation, the copied element, along with its format position in both the main clause and the subordinate clause, is also copied into the CONNECTIVE column, to preserve this fact. However an improved representation using a system which indexes each noun or pronoun will be introduced in the next version of the formatting component. (The indexing mechanism will also be used when copying elements that occur in a conjunction expansion).

Another feature of the information formatting procedure is the treatment of the class of general English operators (time, aspectuals, negation, uncertainty, etc.). In this case, since there is a set of columns for each verb or event in the format, it is necessary to determine what event is being modified in order to map the operator into the particular TIME or MODS subcolumn associated with that event. These operators occur in many different syntactic relations with their operands. They can occur as a verb, operating on the object or subject: *drugs reduced fever, the temperature decreased.* They can occur as adverbs or prepositional phrases: *she was admitted yesterday* or *she had treatment for two weeks.* They can occur as adjectival modifiers, as in *she had frequent transfusions* or *she had a two week treatment.* They can also occur as the head noun in a noun phrase, with the word modified appearing in the left or right adjunct: *she had two weeks of treatment, she had a temperature elevation.* These cases are handled by two general Restriction Language routines. The first checks whether the construction is a noun phrase with the operator as head, and



the event modified in the left or right adjunct. The second routine ($FIND-HOST) handles all the remaining cases. Both routines rely on the structure provided by the parse to locate the relevant constructions. The Restriction Language provides an extremely convenient tool for building such general routines, because it allows complex relations to be built up from the more elementary grammatical routines.

In addition, the parse does not attempt to resolve the question of adjunct placement for certain predictable ambiguities. Therefore the modifier may not appear on its proper host. The parse for the sentence *she was admitted to the hospital for two weeks* will show the time operator *for two weeks* modifying the nearest noun, *hospital.* The formatting transformation must compensate for this by finding the correct operand, namely *admit.* It can do this efficiently during formatting because *hospital,* in the class INST(itution) is one of the classes that does not support TIME or MODS operators. The $FIND-HOST routine requires that the time expression *for two weeks* modify a phrase that has an associated TIME column; if the original host has no associated TIME column, the routine then locates the closest available operand for the time expression, namely the verb *admit* in the above example. Since *admit* is a verb, it does have an associated TIME column, and the time expression *for two weeks* can be mapped into that column.

Time modifiers in particular pose an interesting problem. An analysis of time references has led to the formulation of an EVENT-TIME

TEXT

1. for 2 days before admission
2. at discharge
3. after a month
4. today
5. on Oct. 14, 1971

| EVENT-TIME | | | | |
|---|---|---|---|---|
| TPREP1 | NUM | TIME-UNIT | TPREP2 | REF-POINT |
| for | 2 | days | before | admission |
| | | | at | discharge |
| after | a | month | | |
| | | | | today |
| | | | on | Oct. 14, 1971 |

Fig. 18: Formatted EVENT-TIME expressions

heading, consisting of a direction on a quantified TIME-UNIT expression, in relation to a REF(erence)-POINT. Fig. 18 shows some time expressions in their format representation. A difficulty arises in the representation of the information contained in the REF-POINT. Given the sentence *after hydration, urinalysis normal,* the word *hydration* serves a dual function. On the one hand, it is the REF-POINT in the time expression *after hydration.* On the other hand, if it is entered only in the REF-POINT column, we lose the information that a particular remedy (*hydration* = RX) was administered. This problem has been resolved by adding a second format entry (row) in which the REF-POINT is formatted as a regular event, rather than as part of a time expression. The two format entries are connected by a special relative clause-like connective, EXPAND-REFPT. This treatment is illustrated in Fig. 19.

## 4. *Applications*

The previous sections have described a technique for constructing an information format for a particular sublanguage, and a series of programs for mapping the sentences of sublanguage texts into this representation. The adequacy and utility of the information format as a representation for sublanguage information can be evaluated in terms of computerized information retrieval applications.

### Retrieval Programs

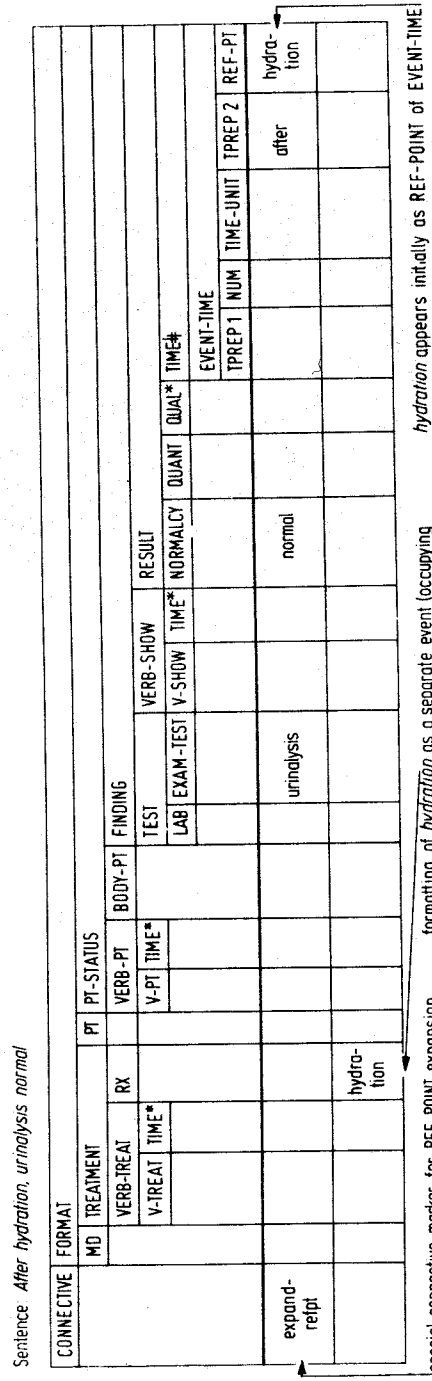
The format representation for the first five sentences of a patient discharge summary is shown in Fig. 20. The accessibility of the data is immediately apparent from inspection of the table. For example, by scanning the SIGN-SYMPTOM (abbreviated S-S) column and the DIAG(nosis)[5] column, we immediately find the patient's problems: dyspnea, effusion, and adenocarcinoma. The BODY-PART column shows what part(s) of the body are affected, namely the lung.

One current experimental application [Sager et al. 1981.] has processed natural language data on earliest symptoms of cancer in the head and neck. A program converts each formatted entry into a code corresponding to the BODY-PART and the SIGN-SYMPTOM. This machine-generated code has been compared to the data as coded by a medical clerk. The machine successfully processed the input data, showing an 86% correlation with the human-generated codes.

A retrieval application for radiology reports has scanned the format representation of the radiology findings on post-operative cancer patients, to determine whether (and when) there was a recurrence of cancer [Hirsch-

[5] In the full format, the diagnosis entries have been placed in a separate column from the SIGN-SYMPTOM entries, to facilitate certain retrieval applications.






*Format columns not shown with their sub-columns for the sake of brevity; MODS (next to TIME) also not shown
*TIME column shown only with its EVENT-TIME sub-column

Fig. 19: Partial information format with REF-POINT of EVENT-TIME expanded into a separate format line.

| CODE | SEQ-NUM | CONN | PATIENT | INST | V-MD | V-TR | RX | BODY-PART |
|---|---|---|---|---|---|---|---|---|
| HIDSH 1 1 1 | 1 | BE PRESENT | | | | | | |
| HIDSH 1 1 1 | 2 | REL-CL | 72 YEAR OLD FEMALE WOMAN FOR THIS | | ADMISSION THE 1ST | | | |
| HIDSH 1 1 1 | 3 | REL-CL | FEMALE WOMAN | | | | | LUNG OF THE |
| HIDSH 1 1 2 | 4 | WHEN | PATIENT THE | | | | | |
| HIDSH 1 1 2 | 5 | WHILE* | FEMALE SHE | | | | | |
| HIDSH 1 1 2 | 6 | | | | | | | |
| HIDSH 1 1 3 | 7 | REL-CL | FEMALE SHE | HOSPITAL A LOCAL | ADMIT TO | | | |
| HIDSH 1 1 3 | 8 | (NTOVO) | | HOSPITAL | FIND | | | |
| HIDSH 1 1 3 | 9 | | FEMALE SHE | HOSPITAL | | | | PLEURAL |
| HIDSH 1 1 4 | 10 | AND | FEMALE SHE | | UNDERGO | | | PLEURAL |
| HIDSH 1 1 4 | 11 | REL-CL | | CYTOLOGY** | | | | |
| HIDSH 1 1 4 | 12 | | | CYTOLOGY | | | | |
| HIDSH 1 1 5 | 13 | REL-CL | FEMALE SHE | | | | P32 | |
| HIDSH 1 1 5 | 14 | | | | | INSTILL INTO | P32 | CAVITY PLEURAL THE LEFT |

NOTE: Each connective (CONN) links the format line on which the connective appears to the next format line(s).
*Second argument of *while (she be in Maryland)* was not formatted, hence does not appear in the table.
**There is no predicate on *cytology* in this line because the sequence *cytology which showed adenocarcinoma* was taken as a separate unit, rather than as a conjoined object of *undergo (she underwent cytology)*.

TEXT: This is the first admission for this 72 year old woman with presumed adenocarcinoma of the lung. The patient was well until 6-74 when she had an episode of severe dyspnea while in Maryland. She was admitted to a local hospital where she was found to have pleural effusion. She underwent pleural biopsy and cytology which showed adenocarcinoma. In 7-74 she had 5 millicuries of P32 instilled in the left pleural cavity.

Fig. 20: The first five sentences of a HISTORY paragraph (from Appendix B) shown in a partial information format.

man and Grishman 1977]. The retrieval program successfully summarized the information in the reports for the eleven patients in the data base. A portion of the computer-generated summary is shown in Fig. 21. This same radiology data base has been used with a question-answering system that allows a user to query the data base via natural language questions [Grishman and Hirschman 1978]. The system can answer questions such as *Did patient 08P003 have an x-ray taken? Was it positive?* and so on.



| V-PT | NEG | MODAL | LAB | NORMALCY | QUANT | S-S | DIAG | TIME | V-TENSE |
|---|---|---|---|---|---|---|---|---|---|
| WITH | | PRESUMED | | | | | ADENOCARCINOMA | | |
| BE | | | | WELL | | | | UNTIL 6/0/74 | PAST |
| HAVE | | | | | SEVERE | DYSPNEA OF | | | PAST |
| HAVE | | | | | | EFFUSION A | | | |
| | | | BIOPSY | | | | | | |
| SHOW | | | | | | | ADENOCARCINOMA | | PAST |
| HAVE | | | | | 5 MILLICURIE PLURAL | | | IN 7/0/74 | PAST |

| Patient | Surgery Date | Reports | Positive Recurrence | Location | Time after Surgery |
|---|---|---|---|---|---|
| 08P003 | 04-17-67 | 8 | YES | Ribs Femoral Pelvis Vertebrae Skull | 1 Year 9 Months 11 Days |
| 08P200 | 09-09-67 | 15 | NONE | Nil | 4 Years 7 Months 3 Days |
| 09P003 | 12-04-61 | 22 | NONE | Nil | 6 Years 11 Months 4 Days |

Fig. 21: Portion of an automatically generated summary from formatted radiology reports.

The Linguistic String Project has also performed several more complex retrieval experiments on the data base created from the eight hospital discharge summaries discussed in the preceding sections. The discharge summaries contain narrative portions (such as the HISTORY paragraph, cf. Appendix B) that are rich in time information. Note, for example, that there are several entries in the TIME column of Fig. 20. In fact, many of the relevant questions about such a data base involve time, e.g., questions such as *Was a blood culture done at admission?* or *Did the chest x-ray show improvement?*, where the expression *improvement* requires a comparison of the initial x-ray with later x-rays. One application in particular has demonstrated the retrievability of time information [Sager et al. 1978]. The retrieval question was drawn from the medical literature, to test a hypothesis that infection precipitated painful crisis in sickle cell patients. The question was: *Did symptoms of infection precede the onset of a painful crisis in patients with sickle cell disease?* Of course, any question concerning possible causal relations will involve time relations between the events, since for event A to be a possible cause of event B, event A must precede event B in time.

Another test of the retrieval has been the application of health care evaluation criteria to the formatted discharge summaries [Hirschman et al. 1979]. The criteria consist of a set of questions about standard procedures performed for a given diagnosis. Discharge summaries that do not meet the criteria are flagged for detailed manual review. Each set of criteria consists of questions concerning diagnosis of the condition (e. g., *Was the throat culture positive for pneumococcus?*), course of treatment (e. g., *Was a chest x-ray taken?*), and outcome (e. g., *Did the chest x-ray show improvement?*). The computer program for application of these criteria produced results in good agreement with those done by a physician reviewer.

### Modularity of Retrieval Routines

Although superficially different, these various applications share a common body of basic routines. This reflects the fact that the different retrieval requests all check relations between the same set of entities, namely the format columns (which correspond to the informational "building blocks" of the sublanguage). For example, the questions *Was the patient afebrile at discharge?* and *Were there any signs of infection prior to onset of painful crisis?* both draw on the FEVER routine (since fever is a sign of infection). Fever has several representations in the format: either as the word *fever* in the SIGN-SYMPTOM column (or one of its synonyms, e. g., *feverish*, *febrile*); or as temperature (in BODY-MEAS(ure) under TEST) above 98.6 degrees (in QUANT). For any mention of *fever* in SIGN-SYMPTOM, it is also necessary to check that it is not negated, i. e., that there is no NEG under FINDING. The negation check is also written as a general routine of



one argument, CHECK-NEG (SLOT) that tests whether the variable SLOT is in the scope of negation. All routines that search for the occurrence of a test, a result, a treatment or any other sublanguage object that can be negated, incorporate the CHECK-NEG routine.

### Normalization

Similarly the time retrieval routines are common to the various applications. However to facilitate the computation of relative time ordering between events, the data base is "normalized" prior to retrieval. Normalization involves making the implicit contextual information explicit in the data base. Since the formatting technique described in section 3 operates on a sentence by sentence basis, any contextually dependent information must be filled in after the sentences have been mapped into the format. This normalization not only facilitates time computation, but also simplifies other retrieval requests as well. Contextually dependent information includes referentials (e. g., in *Swelling of both feet was noted. This progressed and mother brought patient to Pediatric Emergency Service, this* refers to *swelling of both feet*). It includes restoration of an implicitly continued topic, e. g., *Patient was found to have sickle cell disease during 1st admission. Recovery was complete.*, where the topic *patient have sickle cell disease* is carried forward to the second sentence, i. e., *recovery* in the second sentence means *Patient's recovery from sickle cell disease*. It also includes implied progression of time in narrative. For example, in the sequence *She was found to have a pleural effusion. She underwent pleural biopsy which showed adenocarcinoma.*, the time of *biopsy* in the second sentence is taken to be later than the time of *effusion* in the first sentence.

Normalization for the time information consists of assigning to each format row a time in relation to some fixed time point (e.g., date of admission or discharge), or a time in relation to some previously mentioned event in the data base. This requires resolution of referential expressions (found in REF-POINT of EVENT-TIME) such as *prior to the last transfusion*, where the format entry corresponding to the appropriate transfusion must be located. The time normalization program also makes use of change of tense, time information in conjoined sentences, and implicit time progression in narrative [Hirschman 1981]. The normalization routines collect and co-ordinate the various sources of time information, representing it in a uniform manner for each format entry. From this normalized representation, the time relations required by the retrieval requests can then be calculated by a COMPARE-TIME routine. This routine compares the times of two events and indicates either that one event precedes the other or that the two times are equal, or that they overlap, or that there is insufficient information to provide an answer. This last possibility can occur if, for example, the only information available is that both events A and B precede another event C.

## . Conclusion

The retrieval experiments described in section 4 have shown the feasibility of creating a useful data base from natural language input. However these experiments have all involved data bases of limited size, and, particularly in the medical field, significant practical applications require processing data on hundreds of patients (thousands of documents). Our next goal therefore is to refine our procedures and programs to facilitate the handling of large amounts of material with maximum accuracy and minimum human intervention.

The available information about a given sublanguage grows as more sublanguage documents are processed. Much of our present research is directed towards utilizing this accumulation of information. One potential use is a statistically-based treatment of sublanguage selection. In section 1.3, we stated that sublanguage selection was deferred until the formatting stage in our present system. The motivation for this division was to preserve the general character of the first two stages of processing (parsing and English transformational decomposition), that is, to exclude, as much as possible, restrictions specific to a particular sublanguage application. However, if we can accumulate statistically significant numbers of co-occurrence patterns, we believe that it is possible to have the parsing stage draw upon these accumulated patterns, to achieve disambiguation of sentences at a much earlier stage in the processing (which is of course more efficient). The use of a file of possible subject-verb-object or host-modifier patterns to be matched during parsing would also eliminate the need to state each co-occurrence pattern as a specially-written restriction. This approach would still retain the general (non-sublanguage-specific) character of the English grammar, since the selection mechanism would be stated in very general terms, drawing upon a file of co-occurrence patterns specific to the particular sublanguage being processed. It would also simplify the formatting procedure by eliminating the sublanguage-specific processing in this stage. The formatting stage could then be reduced to a much simpler, more general set of rules for mapping words (or phrases) into the format slot corresponding to the subclass of the head of the phrase.

Such a general selectional mechanism would also enable the parser to flag for inspection any new patterns that had not been seen before. This would constitute an important check on the correctness of the processing. Such checks are particularly important in cases where there is omitted information, as for example in the compound noun construction. It is crucial to distinguish the information in the two compound noun phrases *infectious disease consultant* and *infectious disease patient*. In the first case the processing must not misrepresent the information to make it appear that the implied patient has an infectious disease; this would be the effect of simply placing *infectious disease* into the SIGN-SYMPTOM column,



without taking into account the omitted connector between *infectious disease* and *consultant*. On the basis of the set of observed co-occurrence patterns, the missing connection between *patient* and a SIGN-SYMPTOM word (e. g. *infectious disease*) is likely to be V-PT (e. g. *have*: *infectious disease patient* → *patient have infectious disease*). On the other hand, the connection between an MD word *(consultant)* and a SIGN-SYMPTOM *(infectious disease)* is a V-TREAT or a V-MD word (e. g. *treat*: *infectious disease consultant* → *consultant treat infectious disease*, or *study*: *infectious disease consultant* → *consultant study infectious disease*). The selectional mechanism would try to process the compound noun by filling in the missing connection. It would also report that it had seen no prior instances of a SIGN-SYMPTOM + MD compound noun. This would bring the sentence to the attention of the person in charge of checking the processing, to make sure that the representation of the information in the format was correct.

A second potential application of an accumulation of co-occurrence data is the assignment of sublanguage word classes to unknown words, on the basis of their patterns of co-occurrence with known words in the text. Currently all words in a text must be entered into the computer dictionary in order to process the text. Although with each additional text in a sublanguage, there are fewer new words that need to be added to the dictionary, this stage nonetheless creates a bottleneck in processing large numbers of documents. Because the word classes were developed in terms of co-occurrence patterns, the appropriate classification of many of the new words can be deduced from their co-occurrence in syntactic relations with already classified words. For example the word *congestion* is clearly a SIGN-SYMPTOM word in the context *patient complained of chest congestion*. *Congestion* occurs both as the object of a V-PT verb (the object of a V-PT verb is characteristically a SIGN-SYMPTOM word, cf. section 2.1); it also occurs with a BODY-PART modifier *(chest congestion)*, where BODY-PART modifiers occur in general on SIGN-SYMPTOM words, AREA words and LAB words. Thus from the co-occurrence patterns, it should be easy to deduce that *congestion* is a SIGN-SYMPTOM word. In this way, the set of co-occurrence patterns could provide much of the information required to classify new words, thereby shortening the dictionary entry process considerably.

Our approach to the processing of sublanguage texts has been based on established linguistic techniques of distributional analysis. The success of our methods supports the hypothesis that distribution is related to meaning and provides a framework for the study of the information-bearing properties of a sublanguage. On the basis of our successful processing and retrieval experiments in a number of small scale applications, we believe that it will be possible to refine these techniques to create a powerful, general system for large scale automatic processing of natural language documents in a given sublanguage.

*Acknowledgements*

This research was supported in part by National Library of Medicine Grant No. LM02616, awarded by the National Institutes of Health, DHEW, and in part by the National Science Foundation under Grant DSI77-24530 from the Division of Information Science and Technology.

*Bibliography*

[1] ANDERSON, B. B., BROSS, I. D. J., and N. SAGER (1975), "Grammatical Compression in Notes and Records: Analysis and Computation," paper delivered at the 13th Annual Meeting of the Association for Computational Linguistics, Boston; *American Journal of Computational Linguistics 2*, no. 4.

[2] GRISHMAN, R., and L. HIRSCHMAN (1978), "Question Answering from Natural Language Medical Data Bases," *Artificial Intelligence 11*, 25–43.

[3] HIRSCHMAN, L., GRISHMAN, R., and N. SAGER (1975), "Grammatically-Based Automatic Word Class Formation," *Information Processing and Retrieval 11*, 39–57.

[4] HIRSCHMAN, L., and R. GRISHMAN (1977), "Fact Retrieval from Natural Language Medical Records," in *Proc. Second World Conference on Medical Informatics (MEDINFO '77)*, 247–251, IFIP World Conference Series on Medical Informatics Vol. 2, D. B. Shires and H. Wolf, eds., North-Holland, Amsterdam.

[5] HIRSCHMAN, L., STORY, G., MARSH, E., LYMAN, M., and N. SAGER (Oct. 1981) "An Experiment in Automated Health Care Evaluation from Narrative Medical Records," *Computers and Biomedical Research 14*:5, 447–463.

[6] HIRSCHMAN, L., (1981), "Retrieving Time Information from Natural Language Texts" *Information Retrieval Research*, R. N. Oddy, S. E. Robertson, C. J. Van Rijsbergen and P. Williams, eds., Butterworths, London, 154–171.

[7] HOBBS, J., and R. GRISHMAN (1976), "The Automatic Transformational Analysis of English sentences: an Implementation," *International Journal of Computer Mathematics 5*, 267–283.

[8] INSOLIO, C., and N. SAGER (1977), "Parsing Free Narrative: Application to Medical Records," paper presented at the 15th Annual Meeting of the Association for Computational Linguistics, Washington, D. C.

[9] SAGER, N. (1972), "Syntactic Formatting of Scientific Information," Proc. 1972 Fall Joint Computer Conference, *AFIPS Conf. Proc. 41*, 791–800, AFIPS Press, Montvale, N. J. reprinted as chapter 1 of this book.

[10] SAGER, N., and R. GRISHMAN (1975), "The Restriction Language for Computer Grammars of Natural Language," *Communications of the ACM 18*, 390–400.

[11] SAGER, N. (1978), "Natural Language Information Formatting: The Automatic Conversion of Texts to a Structured Data Base," *Advances in Computers 17*, 89–162, M. C. Yovits and M. Rubinoff, eds., Academic Press, New York.

[12] SAGER, N., and L. HIRSCHMAN (1978), "Information Structures in the Language of Science: Theory and Implementation," *String Program Report No. 12*, Linguistic String Project, New York University, New York.

[13] SAGER, N., HIRSCHMAN, L., and M. LYMAN (1978), "Computerized Language Processing for Multiple Use of Narrative Discharge Summaries," *Proc. Second Annual Symposium on Computer Applications in Medical Care*, 330–343, F. H. Orthner, ed., IEEE, New York.

[14] SAGER, N. (1981), *Natural Language Information Processing: A Computer Grammar of English and its Applications*, Addison-Wesley, Reading, Mass.

[15] SAGER, N., BROSS, I. D. J., STORY, G., BASTEDO, P., MARSH, E., and D. SHEDD (1982), "Automatic Encoding of Clinical Narrative," *Computers in Biology and Medicine 12*:1, 43–56.

*Appendix A*

1. Sublanguage Classes Developed for the Clinical Reporting Sublanguage

Note: sublanguage classes generally contain several parts of speech (i. e. nouns, verbs, adjectives, adverbs); however verb classes (classes starting with a V-) contain verbs and nominalized verbs only.

| Class-Name | Description/Examples |
|---|---|
| AMT | amount or degree<br>acute, high, serious |
| AREA | body subpart or area, generally modified by a BODY-PART<br>side, surface, covering, lobe |
| BEH | social behavior<br>work, travel |
| BODY-FUNC | bodily functions<br>eat, walk, flexion, reflex |
| BODY-LOC | location of SIGN-SYMPTOM in relation to BODY-PART<br>encase, radiate, riddle |
| BODY-MEAS | standard body measurement<br>temperature, blood pressure, pulse, respiration rate |
| BODY-PART | part of body or body fluid<br>ankle, endocardial, hair, cerebrospinal fluid |
| DESCR | neutral descriptor of patient state<br>colorless, even, warm |
| DEVEL | physical growth or maturation<br>birth, development, puberty |
| DIAG | diagnosis<br>carcinoma, chicken pox, TB |
| DOCTOR | medical personnel<br>radiologist, consultant, physician |
| EXAM-TEST | test or technique used by doctor during physical exam<br>percussion, palpable, touch |
| FAMILY | members of the family<br>parent, mother, sibling, family |
| INST | institution or clinic<br>emergency room, hospital, cytology |



| Class-Name | Description/Examples |
|---|---|
| LAB | name of laboratory test<br>culture, hematocrit, x-ray |
| LAB-RES | laboratory result (generally of culture)<br>pneumococcal, cloudy, gram negative |
| LOC | location relative to a larger area<br>inferior, left, frontal |
| NORM | normalcy or change towards normalcy<br>well, negative, heal, comfortable |
| OBSERVE | terms of observation (on the part of patient or doctor)<br>mention, note, report |
| PT | patient<br>patient, pt |
| RACE | race<br>Black, Caucasian |
| RECORD | patient record or document<br>chart, history, workup |
| RX | drug, medication or treatment<br>penicillin, transfusion, bed-rest |
| RXFREQ | frequency of administration of medication<br>daily, QD, BID |
| RXMANNER | manner of administration of medication<br>intravenous, oral, aqueous |
| SHAPE | shape or pattern<br>configuration, shadow, silhouette |
| SIGN-SYMPTOM | sign or symptom of abnormal condition<br>ache, coma, degenerate, irritable |
| SURG | surgical procedure<br>appendectomy, operate, incision |
| TESTVIEW | angle or type of x-ray<br>lateral, anterior-posterior, inspiratory |
| TESTYPE | general class of laboratory tests<br>clinical, urinalysis, blood chemistries |
| V-MD | administrative or decision-making activity by doctor<br>admit, care, evaluation, review |

| Class-Name | Description/Examples |
|---|---|
| V-PT | verbs with patient as subject, patient state as object<br>complain, have, sustain |
| V-TEST | verbs with test as subject<br>drop, range, vary |
| V-TREAT | treatment verbs<br>perform, use, inject, give |

2. Connective, Time, Aspectual and Modality Operators

| Class-Name | Description/Examples |
|---|---|
| BEG | words indicating initiation of activity or state<br>form, development, onset, new |
| BECONN | connective stating whole-part or equivalency relation<br>consist of, include, similar to |
| CHANGE | change of state (positive or negative)<br>degenerate, heal, improvement, worse |
| CONN | causal or relational connective<br>associate with, because, lead to, origin of |
| END | termination of a state or event<br>heal, complete, discontinuation |
| EVID | affirmation of existence or occurrence<br>evident, fact, occur, present |
| MODAL | possibility or ability<br>presumed, try, questionable |
| NEG | negation<br>fail, deny, none, not |
| PART | fragment or part<br>drop, element, portion |
| PREP-TIME | time prepositions<br>when, after, until |
| REPT | repetition<br>daily, another, occasionally, recur |
| TM-PER | time period or continuation of a state or event<br>bout, continue, persist |



| Class-Name | Description/Examples |
|---|---|
| TIME-LOC | words indicating location in time<br>later, following, shortly, current |
| TYPE | type or nature<br>type, kind, sort |

*Appendix B*

Page 1 of a 2-Page Hospital Discharge Summary

Discharge Summary
ADMISSION DATE: 9–6–xx
DISCHARGE DATE: 10–9–xx

HISTORY:

This is the first admission for this 72 year old woman with presumed adenocarcinoma of the lung. The patient was well until 6-xx when she had an episode of severe dyspnea while in Maryland. She was admitted to a local hospital where she was found to have carcinoma. She underwent pleural biopsy and cytology which showed adenocarcinoma. In 7-xx, she had 5 millicurries of P32 instilled into the left pleural cavity. Workup also included a liver scan which was normal, normal upper GI and cholecystogram, normal IVP, mammograms, bone scan and pancreatic scan. The patient apparently underwent a complete laparotomy 2 weeks prior to her first clinic visit. She has been followed in the oncology clinic since Aug. 9. Review of slides showed malignant cells suspicious of bronchial malignancy. Patient is presently admitted for chest tube insertion and possible pleurectomy and thoracotomy because of recurrent nature of effusion. For the past week, the patient has had increasing shortness of breath on exertion, and now even while at rest. Patient has had a cough for the past 4–5 months which has increased in the past few weeks. This occurs in the morning without sputum production. She smoked 1–1½ packs of cigarettes per day for 40–50 years and stopped in Dec. 1972. She has noted increasing fatigue over the past few weeks.

PHYSICAL:

Blood pressure 120/80; pulse 100; respirations 28; temperature 37.

HEENT: ears packed with cerumin, otherwise within normal limits.

NECK: supple, without nodules or thyromegaly.

| | |
|---|---|
| CHEST: | no expansion on left; right lung clear to percussion and auscultation; left, decreased breath sounds to 3/4 way up, decreased vocal fremitus; chest tube in place, rales present at left base. |
| HEART: | PMI 5th intercostal space. |
| ABDOMEN: | midline scar, no masses or organomegaly, liver 11 cm by percussion. |
| RECTAL: | no masses, guaiac negative. |
| EXTREMITIES: | questionable deep tenderness on palpation on left lower extremity, otherwise normal. |
| NEUROLOGICAL: | grossly intact, except for decreased Babinski. |

BIOCHEMISTRY:

CA 10.2; Phos 3.9; BUN 14; uric acid 6.8; total protein 7.6; albumin 4.1; bilirubin .83; alk phos 133; LDH 190; SGOT 22; Glucose 162; PT 13.5; Creat 1.0, CO2 28; Na 146; K 4.3; Cl 99.

HEMATOLOGY: WBC 13; RBC 4.5; HGB 12; HCT 37; Platelets 510,000.

URINALYSIS: Specific gravity 1.001; pH 5; WBC – many. Sq ep – occ.

*Appendix C*

Output of the Parsing Program
Parse tree of 2nd sentence from the HISTORY paragraph (Appendix B)

Explanation of Node Name abbreviations

| | |
|---|---|
| TEXTLET, OLD-SENTENCE, MORESENT | nodes to allow more than one CENTER within a sentence |
| CENTER | independent unit terminated by an ENDMARK |
| INTRODUCER | slot for paragraph headings (e. g., *HISTORY*:) |

*Terminal symbols*

| | |
|---|---|
| N | noun |
| ADJ | adjective |
| TV | tensed verb |
| T | article |
| PRO | pronoun |
| CS0,CS1 | subordinate conjunctions |
| P | preposition |



| | |
|---|---|
| V | verb |
| Q | quantifier |

*Phrase types*

| | |
|---|---|
| TENSE | slot for modals (e. g., *will*, *may*) |
| SA | sentence adjunct |
| OBJECTBE | object set of the verb *be*, including participles |
| OBJBE | non-verb containing objects of *be* |
| NSTG | noun string |
| NSTGO | noun string in object |
| ASTG | predicate adjective string |
| CSSTG | subordinate conjunction string |
| LNR | noun with left and right adjuncts |
| LAR | adjective with left and right adjuncts |
| LAR1 | prenominal adjective with left and right adjuncts |
| LDATER | date with left and right adjuncts |
| LTR | article with left and right adjuncts |

*Left adjuncts*

| | |
|---|---|
| LN | left adjuncts of noun |
| LV | left adjuncts of verb |
| LA | left adjuncts of adjective |
| LCDA | left adjunct position of compound adjective |
| LCS | left adjunct of subordinate conjunction |
| LT | left adjunct of article |
| LDATE | left adjunct of date |
| LP | left adjunct of preposition |

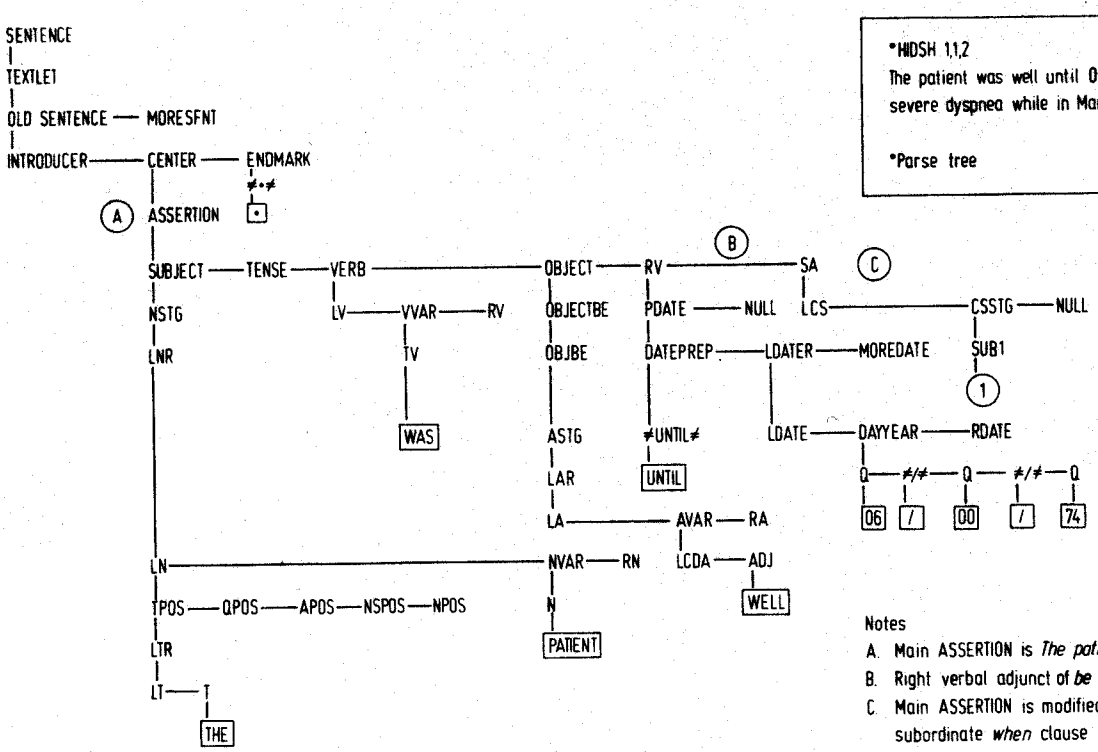
*Right adjuncts* – similar to left adjuncts

*Local variants*

| | |
|---|---|
| NVAR | variants of noun |
| AVAR | variants of adjective |
| VVAR | variants of verb |

*Left noun adjunct (LN) positions*

| | |
|---|---|
| TPOS | article position |
| QPOS | quantifier position |
| APOS | adjective position |
| NSPOS | possessive-of-type position |
| NPOS | noun position |



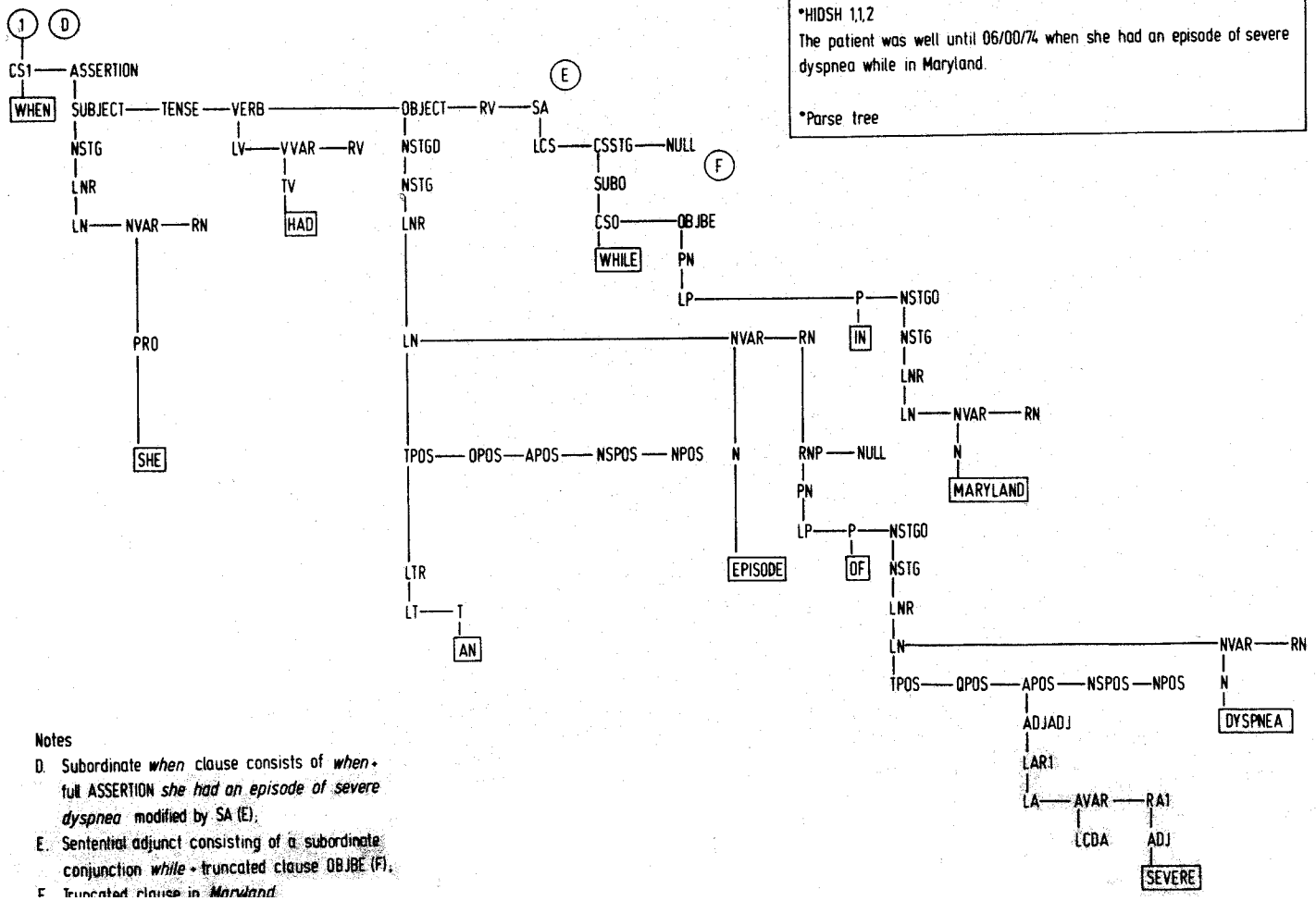

*HIDSH 1,1,2
The patient was well until 06/00/74 when she had an episode of severe dyspnea while in Maryland.

*Parse tree

Notes

A. Main ASSERTION is *The patient was well*, with right verbal adjunct (B);
B. Right verbal adjunct of *be* is *until 6–74*,
C. Main ASSERTION is modified by a sentential adjunct consisting of a subordinate *when* clause (see next page).



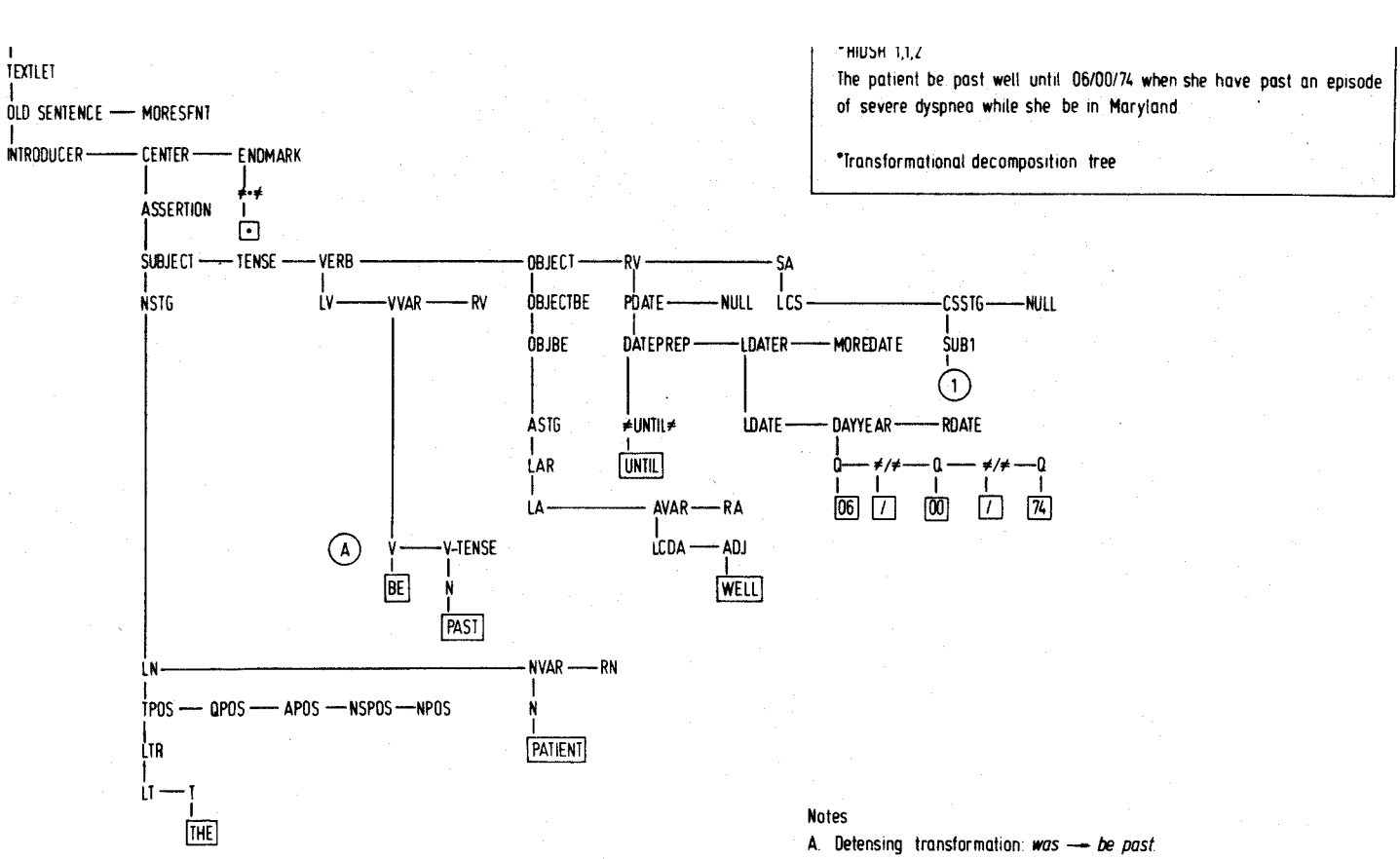



*HIDSH 1,1,2
The patient was well until 06/00/74 when she had an episode of severe dyspnea while in Maryland.

*Parse tree

Notes

D. Subordinate *when* clause consists of *when* + full ASSERTION *she had an episode of severe dyspnea* modified by SA (E);
E. Sentential adjunct consisting of a subordinate conjunction *while* + truncated clause OBJBE (F);
F. Truncated clause in *Maryland*.

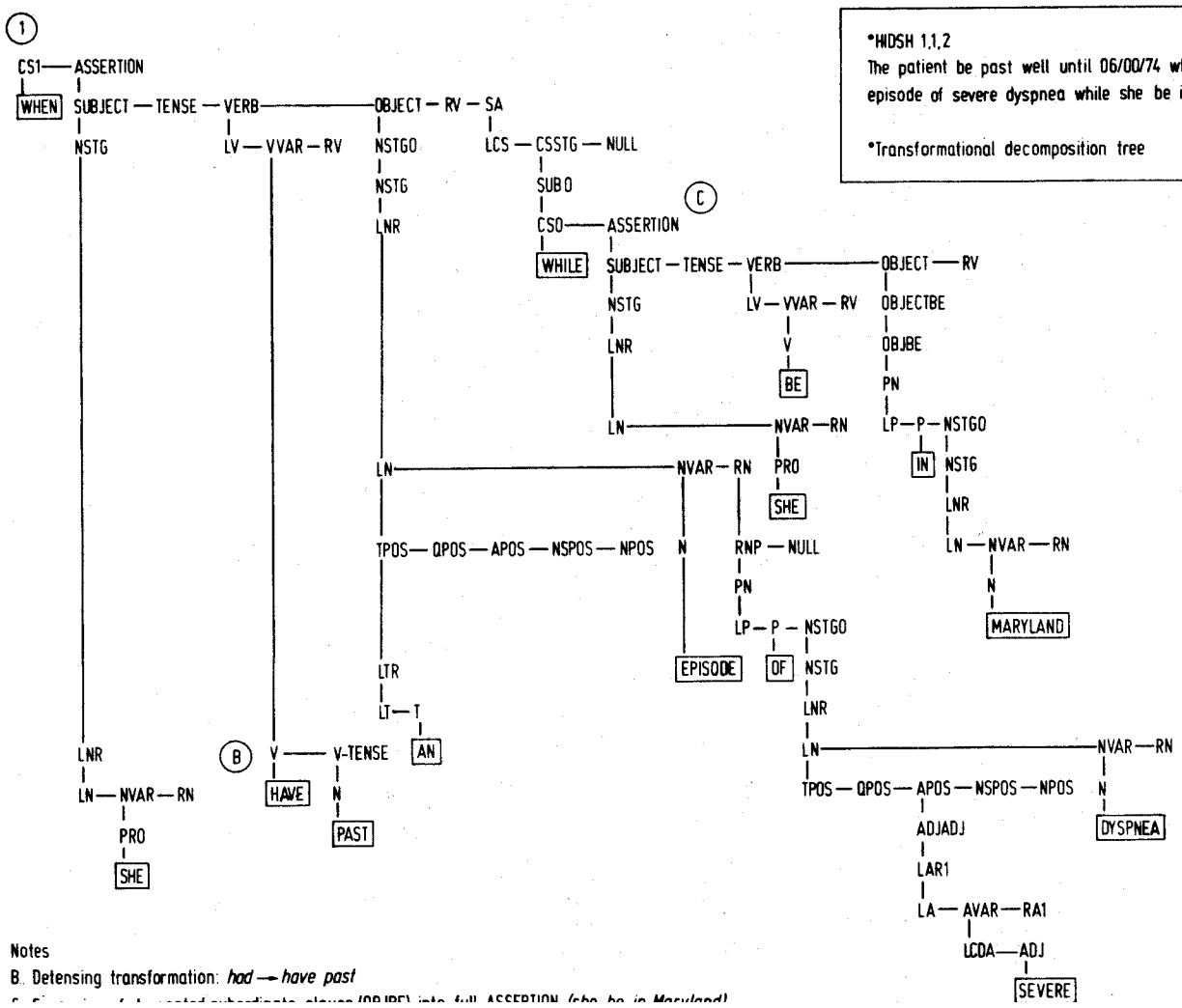



*HIDSH 1,1,2
The patient be past well until 06/00/74 when she have past an episode of severe dyspnea while she be in Maryland

*Transformational decomposition tree

Notes
A. Detensing transformation: *was* → *be past*






*HIDSH 1,1,2
The patient be past well until 06/00/74 when she have past an episode of severe dyspnea while she be in Maryland

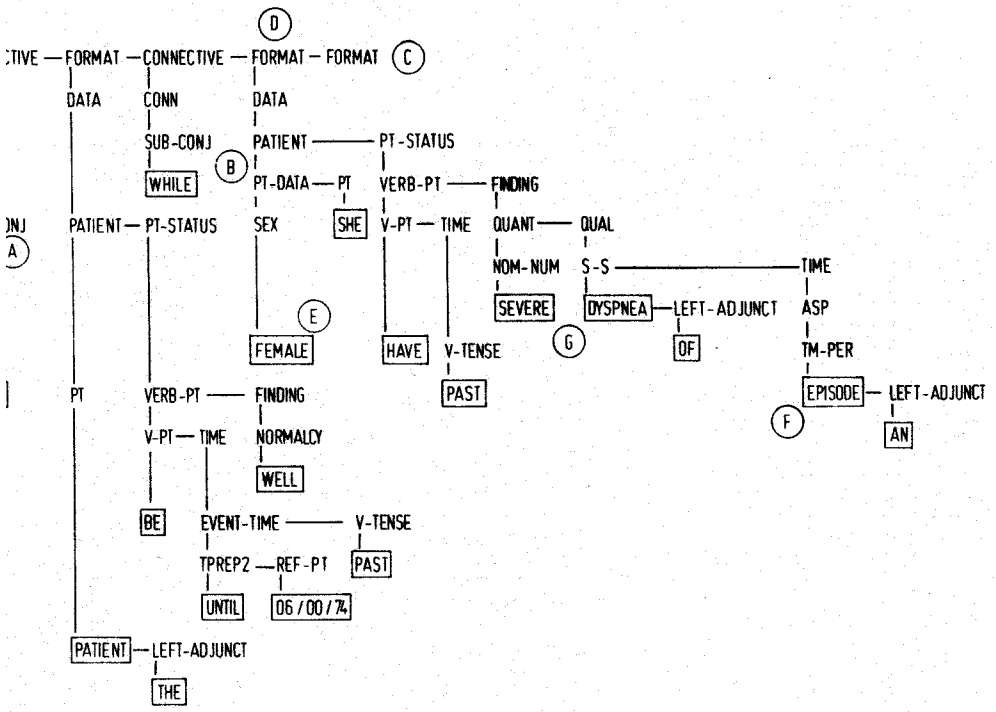
*Transformational decomposition tree

Notes
B. Detensing transformation: *had* → *have past*

*Appendix E*

## Format (represented as a tree) for 2nd sentence in HISTORY paragraph

1,1,2
patient was well until 06/00/74 when she had an episode of severe dyspnea while in Maryland.
overs % she be in Maryland.




non-empty nodes appear in the tree;
s appearing in the format are enclosed in boxes, to distinguish them from the headings of the format.
nd right adjuncts format entries are labelled in the tree.

st connective *when* connects first FORMAT (corresponding to the main clause) with the *when* clause;
cond connective *while* connects the second FORMAT (corresponding to the *when* clause *she had an episode of severe spnea*) with the third FORMAT (empty) corresponding to the *while* clause;
ird format is empty because the statement *she be in Maryland* does not fit into clinical reporting sublanguage;
aterial from the *while* clause is not formatted, i. e. "leftover"; the leftovers flag unformatted material;
*male* is filled from the feminine pronoun *she* under SEX in PATIENT-DATA;
*isode* is a TIME-PERIOD operator on *dyspnea*;
*were* is a NON-NUM(erical) QUANT(ifier) on the QUAL(itative) finding *dyspnea*.

# Contents