From text to structured information—Automatic processing of medical reports*

by LYNETTE HIRSCHMAN, RALPH GRISHMAN and NAOMI SAGER
New York University
New York, New York

ABSTRACT

This paper describes the analysis and processing programs for a set of natural language texts in a medical area (x-ray reports on patients with breast cancer). The programs convert the information in the text into a tabular form suitable for further automatic information processing (e.g., editing of records, question answering on the data collected, or statistical summaries of the data). To set up a tabular form appropriate for the data, we first perform a manual linguistic analysis on a sample of the texts. From this we obtain the word classes and the form of the table (called an information format) for this type of material. We then apply the series of processing programs to the sentences of the texts. Each sentence is parsed with the Linguistic String Parser English grammar in order to obtain its grammatical structure; certain standard English transformations are then applied to regularize the grammatical form of the sentence; and finally a set of "formatting transformations" map the words of the sentence into the slots of the format or table, in such a way that the sentence is reconstructible (up to paraphrase) from its representation in the table. The results of applying these programs to a corpus are described. This procedure enables us to convert a natural language corpus into a structured data base.

INTRODUCTION

An essential part of the effective management of scientific and technical information is the efficient retrieval of information from a large body of text. One example of this is the retrieval of documents from a large collection of scientific articles, in response to a user's request. Another example of the same problem is the extraction of data from medical reports for statistical purposes, or for fact retrieval.

The key to efficient retrieval lies in the appropriate structuring of the information. For document retrieval, this may involve the extraction of key terms for each document. For medical records, it may involve transferring the most essential information into separate tables. These tasks pose a considerable burden on the preparer of the document. In addition, each such structuring will be appropriate only for the retrieval of certain types of information from the data base.

What is required therefore is a procedure for the automatic structuring of the natural language material itself, in such a way that all the information is preserved. The Linguistic String Project of New York University has been engaged in a long-term effort to develop techniques for processing textual information. These techniques are based on distributional analysis and computerized parsing of English texts. We intend in this paper to give an overview of our approach and to describe briefly our latest experiments.

OVERVIEW

Since we are dealing with textual data, structuring the information means, first of all, structuring the sentences. The question then is: what sort of structure should be assigned to the sentences? One alternative is some kind of surface parse tree. PROTOSYNTHEX I,[1] one of the earliest systems for information retrieval from natural language texts, attempted to use dependency analysis to match requests for information with sentences in the data base. However, surface analysis alone is inadequate for such information processing; one limitation is that it does not take into account possible differences between data and request due to grammatical paraphrase. For ex-

* This investigation was supported in part by research grant 1-RO1-LM-02616 from the National Library of Medicine, National Institutes of Health, DHEW, in part by Public Health Service Research Grant No. CA-11531 from the National Cancer Institute, and in part by research grant SIS75-22945 from the National Science Foundation, Office of Science Information Service. Development of the program for transformational analysis was supported in part under Contract No. NOOO14-67A-0467-0032 with the Office of Naval Research.

ample it would fail to match a request stated as an active sentence with an otherwise identical sentence in the data base which was in the passive voice.

It has long been recognized[2] that the effects of such paraphrastic variation can be overcome by performing some type of transformational analysis on the sentence. Transformational decomposition, following the theory of Harris, or deep structures, following the theory of Chomsky, can be used to reduce grammatical paraphrases to a standard form. A Linguistic String Project study in 1970 showed that Harrisian transformational decompositions could be useful in matching technical articles with information requests.[3] Such techniques can be used to structure a variety of texts; however, the resultant structures provide only general grammatical relations (subject, object), which are not directly related to the semantic or informational classes in a specific scientific subfield. In other words, the categories of English grammar are too general for information structuring.

It is possible to write a grammar specific to the use of language in a particular subfield of science, employing the same methods used to write descriptive grammars of whole languages. The resulting sublanguage grammar yields structures suitable for information processing: the word classes of this grammar are the word classes of semantic interest in the subfield; the overall arrangement of classes provides a format for the information content of subfield text sentences. For example, the grammatical structure of medical reports includes categories for patient, type of test, body organ tested, date of test, etc. Such an organization can greatly facilitate information retrieval or statistical manipulation of the data. On the other hand, each scientific field and type of text has its own structure. This means that a detailed linguistic analysis is required every time a new class of text is to be handled.

In this paper we describe an experiment in the computer formatting of material from medical records. Our previous papers have described the method of sublanguage analysis and information formatting for more complex textual material,[4,5] as well as the battery of programs which have been developed for text processing.[6,7] Here we focus on the problem of mapping text sentences into information formats. In the sections which follow, we will describe how the format for a particular type of medical narrative was derived, and how sentences are automatically transformed into structured information, as specified by the format. We will also indicate how the process of deriving formats may be automated or partially automated, and how the structured information of the formatted sentences can be used.

THE TEXTS

For our initial experiment in the computer formatting of texts, we chose to work on medical records. A set of follow-up reports on patients with cancer was provided to us in machine readable form, as part of a collaborative research project with Dr. I. D. J. Bross of Roswell Park Memorial Institute. The reports included laboratory tests, pathology reports, radiology reports, records of treatment, and discussion of medical problems. A linguistic analysis of some of these reports was done at Roswell Park.[8] We chose to process one particular type of report, identified as "Findings R(adiology)." This material was selected because it contained both full sentences and sentence fragments, a combination typical of the compressed note-taking style of much medical narrative (e.g., *x-rays not taken,* or *nothing to indicate metastasis*). The limited vocabulary of Findings R and the frequent paraphrasing of the various types of medical information made it possible to define valid word classes and formats on a limited corpus.

The corpus consisted of 159 Findings R reports on 11 patients, containing a total of 188 sentential units.* Due to frequent repetition of certain formulaic expressions, such as *x-rays negative,* only about half of the sentential units (86/188) were distinct, ignoring differences in date.

CREATION OF SUB-LANGUAGE FORMATS

To convert the medical information contained in a sentence into tabular form, we create a table (or format) with slots for each class of relevant information. The definition of a set of formats for a particular sub-field is done in two steps: first we perform a distributional analysis on the parsed sentences to obtain the sub-language word-classes; we then use the distribution of the word-classes to define the formats.

Distributional analysis involves classifying together words which occur in the same syntactically defined environments; for example verbs which occur with the same subjects and objects would form a word class. We begin building each class by finding a few words which occur frequently and share a number of environments. These words form the "core" of the new class. We then enumerate the environments in which these core words occur, and look for other words which share some of the same environments. If these other words occur primarily in the same environments as the core words, we add them to the class. This process can be illustrated with the NTEST (Noun TEST) class. The words *x-ray* and *film* share many environments, and are thus selected as the core of a new class. The characteristic environments in which they occur are:

* A sentential unit is a word sequence ending in a period; a sentential unit may contain more than one sentence or sentence fragment: *chest x-ray unchanged, nothing to indicate metastatic disease.*

(1) [chest] {x-ray(s) / film(s)} [RN] [show / --] [LN] {change(s) / metastasis / metastases} [RN]

(2) [chest] {x-ray(s) / film(s)} [RN] [be / --] negative [RN]

Here braces enclose alternative elements and brackets enclose optional elements; LN and RN designate left and right adjuncts (modifiers) of the noun. Note that the dash is treated as a word of the sentence. Looking for other words which appear in these environments, we find:

(a) *Mammograms* no change. . . .
(b) Metastatic *series* showed extensive osteolytic metastases. . . .
(c) Metastatic bone *survey* -- negative.
(d) Flat *plate* -- mild degenerative changes. . . .
(e) Flat *plate* of abdomen -- shows lumbar spine to be riddled with multiple metastatic areas.

Since the environments of *mammograms, series, survey,* and *plate* match either environment (1) or (2) of the core words, they are added to the NTEST class.

We have implemented this approach to word classification in a computer program, although using a somewhat different procedure than that described above.[9] The program has been applied to the Findings R data and to other texts. Both the manual and computerized methods successfully classify all of the frequent words and some of the infrequent words.

To capture more of the infrequent words, we use a second-order distribution analysis procedure. In the characteristic environments of each class, we replace each word which has already been classified, by the name of its class. Consider the two environments of *x-ray* and *film* given above. At this point *chest* has been assigned to the NBODY class; *x-ray* and *film* to the NTEST class; *show* to the VSHOW class; *be* to the VBE class; *change* to the NCHANGE class; *metastasis* (*-ses*) to the NCONDITION class; and *negative* to the NONPATHADJ (non-pathological adjective) class. Replacing each word in the environments listed above by its class name, we get:

(3) [NBODY] NTEST [RN] [VSHOW / --] [LN] {NCONDITION / NCHANGE} [RN]

(4) [NBODY] NTEST [RN] [VBE / --] NONPATHADJ [RN]

In similar fashion, we take the environments of each unclassified word and replace the words by class names where possible. For instance, there are two occurrences of *scan*:

(f) The liver scan was normal.
(g) Brain scan shows midline lesion.

Replacing words by class names, we obtain:

(h) NBODY scan VBE NONPATHADJ
(i) NBODY scan VSHOW LN CONDITION

Since these two sentences match the environment for NTEST words, we add *scan* to the NTEST class.

There are some words which occur so infrequently (once or twice in the corpus) that we cannot rely on distributional analysis to classify them. However these words must be assigned to a sublanguage class if the sentences in which they occur are to be correctly formatted. (Words are assigned to format slots on the basis of membership in a word class.) In these cases we either extend the criteria of a sub-class in reasonable ways, or if all else fails we use our knowledge of the meaning of a word to fit it into a subclass. On this basis we add to the NTEST class the words *auscultation, percussion, urinalysis,* and *view,* each of which occurred only once in the corpus.

Once we have defined the sublanguage word classes, we can use the word classes to define the sublanguage formats. A format is constructed so that:

1. equivalent pieces of information in different sentences will map into the same format slots;
2. each informationally significant word in a sentence is mapped into a separate slot of the format;
3. in each sentence, certain slots of the format may be empty, if the sentence does not contain that particular type of information;
4. not every word in a sentence will receive its own format slot: certain modifiers (e.g., *the*) are simply left as adjuncts on their head noun, if they contain no sublanguage information, or if they never occur independently of a particular word class;
5. *all* the words of the sentence are mapped into the format, preserving their original order of occurrence, with the exception of certain allowable paraphrastic permutations.

Once the sentences are formatted, we know exactly where (what slot or slots) to check, in order to find any particular type of information, in any sentence. However the formatted sentence will resemble the original unformatted sentence very closely, since no words are lost, and word order is preserved up to paraphrase. It is surprising that the sublanguage sentences are so highly structured that an information format can be constructed in this way, but it is just this structure that makes it feasible to do natural language processing on these texts.

We begin to build the format by taking a sentence of the corpus:

(a) Chest x-ray 12-6 shows no evidence of metastasis.

We replace the words by their sublanguage classes:

(b) [NBODY NTEST DATE] [VSHOW]
chest x-ray 12-6 shows
subj verb
[NEGATIVE NSHOW P NCONDITION]
no evidence of metastasis
obj

Each significant word gets its own slot. Of these words, only P (preposition) has no significance beyond its role as syntactic marker; it is therefore included as an adjunct of NCONDITION. We can now write our first tentative format. The format slots are given names related to the type of information they will contain. The gross syntactic structure of the parsed sentence provides some additional groupings of the format slots into TEST (subject) and FINDING (predicate). In Table I words in () are adjoined to the main word in the slot.

Next we take another sentence and again replace the words by their word classes:

(c) [DATE NTEST P ADJSPINE NBODY]
10-26 film of lumbar spine
subj
-- [NEGATIVE NCHANGE]
no change
-- obj

The subject of sentence (c) contains the word classes DATE, NTEST, NBODY, but in a different order than sentence (b). However there are paraphrastic transformational relations that allow the date (a time expression) to be on either side of the subject; and a paraphrastic transformational relation between the two noun phrases:

N_1 N_2 $\leftrightarrow$ N_2 of N_1
chest film film of chest

Since the subjects of sentences (b) and (c) contain the same kinds of information, this information must be mapped into the same format slots in both cases. Changing the word order of sentence (c) for formatting is permissible, since it only involves paraphrastic permutations. Therefore *10-26*, *film*, and *spine* map into the format slots TESTDATE, TESTN, and TESTLOC as set up in format #1. *Lumbar* is not assigned a format slot of its own, but is left as an adjunct on *spine*, because ADJSPINE adjectives occur only on *spine* in this corpus; that is, they have no independent status and do not get a separate column of the format.

Next we must decide what to do with the symbol "--" which appears between subject and object in sentence (c). Should it be assigned to a new format slot, or can it be mapped into the VERB category? If we examine its distribution, we find that it has the distribution of VSHOW in certain cases, and of VBE in others, e.g., *Chest x-ray -- no evidence* and *Chest x-ray -- negative*. It is therefore appropriate to map it into the VERB slot. Finally, we must decide where to put NCHANGE in the format. Its distribution differs from NSHOW and NCONDITION; in particular it can occur in the same sentence with words from these two classes:

(d) No evidence of recurrence
NSHOW NCHANGE

(e) No callus formation
NCONDITION NCHANGE

Clearly the class NCHANGE is not in complementary distribution with either NSHOW or NCONDITION. We must create a new slot in the format between INDICATION and MED-FINDING to house it. Our revised format #2 is shown in Table II with formatted sentences (b)-(e):

In this manner we build up the format on the basis of a limited number of sentences. The adequacy of the format created can be tested by using it in the formatting of a different set of texts. The x-ray format made up from part of the Findings R data has been tested both against other Findings R data and against a different set of x-ray data from patients with sickle cell disease. In both cases it was found adequate to format the radiology material.

TABLE I

FORMAT #1

TEST			FINDING			
TESTLOC	TESTN	TESTDATE	VERB	NEG	INDICATION	MED-FINDING
NBODYPT	NTEST	DATE		NEGATIVE	NSHOW	NCONDITION
chest	x-ray	12-6	shows	no	evidence	(of)metastasis

TABLE II

FORMAT #2

	TEST			FINDING				
	TESTLOC	TESTN	TESTDATE	VERB	NEG	INDICATION	CHANGE	MED-FINDING
	NBODYPT	NTEST	DATE	VSHOW --	NEGATIVE	NSHOW	NCHANGE	NCONDITION
b)	chest	x-ray	12-6	shows	no	evidence		(of)metastasis
c)	(of)(lumbar) spine	film	10-26	--	no		change	
d)					no	evidence	(of) recurrence	
e)					no		formation	callus

Not all the entries in Findings R report the results of a test. There are a few sentences that refer directly to the patient:

(g) Patient given penicillin for 9 days.

(h) Patient to return in one month for repeat x-ray.

Clearly these sentences require a different format from the one being developed above. Since there are so few sentences of this type, a much larger corpus would be required to define a format for sentences (g) and (h), but as these sentences illustrate, even in a restricted subfield of a medical report, several formats may be needed to represent the different types of information encountered.

FORMATTING THE TEXT

Once the format is defined, the sentences must be mapped into the format. As before, it is important to have a procedure which can be generalized to texts in other subfields. Our procedure is built around the Linguistic String Parser, a powerful system for language analysis which provides the mechanism for parsing sentences with a context-free grammar augmented by restrictions;[7] it also provides the machinery for performing transformations on parsed sentences,[10] and a higher level language (the Restriction Language) for writing restrictions and transformations.[11]

Sentence formatting is done in three stages:

1. determination of sentence structure by linguistic string analysis;
2. regularization of certain sentence structures by use of general English transformations;
3. mapping of transformed parsed sentences into format slots, using specialized "formatting transformations."

We will briefly consider each stage in turn.

Linguistic string analysis

Linguistic string analysis provides a structural description of the sentence in terms of a specified set of linguistic strings. The assignment of a word to a format slot depends on its role in the sentence structure, as well as on its word class, so that a determination of sentence structure is a prerequisite to formatting. For example syntactic analysis resolves part-of-speech ambiguities, so that the word *left* is identified as a verb in

(a) The patient left the hospital.

but as an adjective in

(b) X-ray of the left lung showed metastasis.

The LSP string grammar was originally designed to handle only complete English sentences; it provides a broad coverage of English syntactic constructions and together with its associated word dictionary, has been used to analyze English scientific texts. In order to process the note-style and incomplete sentences of the medical reports, we made four changes in the grammar.

First, the grammar was expanded to handle the sentence fragments by adding a small number of new productions to the context-free component. Five types of fragments were allowed.

1. A sentence with subject and object but either without verb or with a dash (--) in place of a verb: *Chest x-ray -- no change., 10-6 x-ray negative.*
2. An adjective with its adjuncts: *Negative for metastatic disease.*
3. A noun with its adjuncts: *No evidence of change.*
4. A passive sentence without subject or *be*: *Not done on previous exam.*
5. A sentence preceded by a noun phrase. *Chest x-ray 4-6-71 chest film shows no evidence of fluid.*

Second, one restriction in the grammar was removed in order to accommodate the note-taking style of the text: this was the count noun restriction, which requires that a singular count noun have an article or some other appropriate form of modifier before the noun. For example, the Findings R text contains sentences like *X-ray shows lesion,* whereas in normal English both *x-ray* and *lesion* must be preceded by an article: *The x-ray shows a lesion.*

Third, certain constructions that were unlikely to occur in this type of text were eliminated from the grammar, for example, the question constructions. This pruning of the grammar speeded up the sentence analysis considerably.

These first three changes were designed to accommodate texts in a note-taking style and would be applicable to any subject area. A fourth change, needed to handle certain types of ambiguity, required the use of word classes and selectional restrictions specific to the sublanguage grammar of radiology reports.

One such type of ambiguity is a predictable structural ambiguity, which must be resolved in order to format the sentences correctly. This type of ambiguity can arise from modifiers on conjoined material. For instance, the sentence

(c) X-rays of lumbar spine and chest showed lesions.

may be analyzed as any one of the following:

(d) X-rays of lumbar spine showed lesions and chest showed lesions.

(e) X-rays of lumbar spine showed lesions and x-rays of lumbar chest showed lesions.

(f) X-rays of lumbar spine showed lesions and x-rays of chest showed lesions.

Such ambiguity is inherent in the syntactic construction, and has nothing to do with the particular words involved. Only sublanguage selectional restrictions can resolve it. In this example, *lumbar* is an ADJ-

SPINE which modifies only the noun *spine*, eliminating reading (e). Since *spine* and *chest* are both NBODY, it is more likely that they are conjoined than words of different classes (e.g., *x-ray*, an NTEST and *spine*). This eliminates reading (d) leaving the correct reading (f).

Another type of ambiguity arises from an "overrich" lexicon--a lexicon for all of English, containing possible uses of words that would never occur in this sublanguage. The sentence

(g) No report of x-rays being taken.

received a parse in which *being* was taken to be a noun (as in *human being*) which was the object of a missing verb *be* or *show* derived from a sentence:

(h) No report of x-rays shows a being which has been taken.

There were several ways to deal with this kind of ambiguity. We could have used selectional restrictions on the subject and object of *taken*, or we could have placed tighter restrictions on the construction with an omitted verb. However we chose what seemed the simplest and most direct approach for such cases: we created a special x-ray dictionary, by editing the general English dictionary to remove word classifications (e.g., *being* as a noun) which would never occur in the Findings R text.

English transformations

In the Linguistic String Parser, the transformations are applied to the output of the string analysis. The function of the transformations is to regularize the parse trees, reducing the variety and complexity of structures present. For example, the sentences

(i) X-rays of chest and pelvis negative.
(j) X-rays of chest negative and x-rays of pelvis negative.

contain the same information. By transforming the parse tree for the first sentence into the tree for the second sentence we produce a more regular set of structures in which only full sentences (or sentence fragments) are conjoined. We also transform relative clauses into complete sentences; for example, we would convert

(k) X-rays showed a lesion which may be metastatic.

to

(l) X-rays showed a lesion such that the lesion may be metastatic.

The gain achieved in performing this transformation is that the complete sentences derived from relative clauses can then be formatted in the same way as any other sentence; no special process for formatting relative clauses is required.

These transformations are written for all of English; they do not make use of any information specific to the Findings R sublanguage. There are a large number of English transformations, but only a very few have been used in this application. This is because transformations expand compressed material into a more regular form by filling in certain pieces of redundant information, or information retrievable from context (like the verbs *be* or *show*). If a particular type of information is always omitted in a certain class of texts, no regularization is achieved by trying to fill in this missing information. For example, the word *x-ray* can be used both as a verb and as a noun, so we could have an English transformation to convert sentences with the noun to sentences with the verb; in Findings R, however, *x-ray* is used only as a noun, and no regularity would be gained. Moreover, the verb requires a subject--the taker of the x-ray--which is never present in this text. As a result, the only two English transformations used are the conjunction expansion and relative clause expansion described above. However, in more syntactically complex material or less abbreviated material, there might be a real benefit from a greater regularization of the syntax (via transformations) before attempting to format it.

Formatting transformations

The formatting transformations transfer the words from the parsed sentences to the appropriate slots in the format. They use the same transformational mechanisms built into the Linguistic String Parser to handle the English transformations. Because these mechanisms are set up to map trees into trees, the format is first created as a tree; after it has been built, it can be written out in the tabular form shown in Tables I-III. Formatting transformations move the words from the original ASSERTION or FRAGMENT node in the parse tree into the format slots. As a result, at the end of the formatting process, the FORMAT has the words of the sentence in it, while the original ASSERTION or FRAGMENT node is empty. This provides a check on the completeness of the formatting process.

Three kinds of transformations can be distinguished. The first type of transformation sets up the format slots under the node FORMAT. For sentences which contain a verb or adjective connecting two findings or pieces of data (e.g., *related to, compatible with, typical of*), the format is augmented with a CONNECTIVE slot and an additional set of FORMAT slots. It is necessary to add this new FORMAT to provide an empty set of slots for the second finding. Relative clauses are treated similarly: a CONNECTIVE slot is added, with a relative clause marker placed in the CONN slot under CONNECTIVE; and the assertion contained in the expanded relative clause is mapped into the new set of format slots.

Once the format slots have been set up, the remaining transformations each map words of one class into the appropriate format slot. These transformations fall into two groups. One type requires little if any syntactic or co-occurrence information; it simply searches the parse of the sentence for a word having both a particular syntactic category and a certain sublanguage word class, and then maps the word into the format slot associated with that word class. For example, the T-NTEST transformation looks for a NOUN of class NTEST. When it finds such a word in the sentence being formatted, it moves the word, together with its adjuncts, into the appropriate format slot (TESTN).

A different type of formatting transformation is required for a word that can go into one of several slots depending on what it modifies (e.g., negative and indefinite words). This class of transformations relies heavily on the availability of syntactic information from the parse. For example, *not* can go in any one of three slots, depending on what kind of verb it negates. Therefore the T-NOT transformation must apply before any of the verb transformations have moved the verb into its format slot: co-occurrence relations must be checked in the parse tree, where the syntactic relations are still explicit. Once the verb has been moved into the format, the syntactic relations have been translated into informational relationships and are no longer explicitly expressed. When the T-NOT transformation finds a *not*, it checks the main verb occurring in the same string with *not*. If the verb is VSHOW (e.g., *X-rays do not show metastasis.*), the *not* goes into NEG in FINDING. If the verb is VDONE (as in *not done*) the *not* goes into the NO-TEST slot, under TEST, because the class VDONE occurs only with NTEST nouns; if *not* occurs with VDONE, it necessarily negates the existence of a test, even if no NTEST word occurs in the sentence. Finally if the *not* negates a word that connects two findings (e.g., *is not related to, is not compatible with*) it will go into the NEG-CONN slot under CONNECTIVE.

The set of formatting transformations can be viewed as a set of special sublanguage transformations which reduce various sublanguage paraphrases to a standard representation in the format. For example, to find out if a test was performed, we need only inspect the NO-TEST column of the format. If it is empty, a test has been performed and we can find the type of test by looking in the NTEST slot. Or if we want to know when the first abnormality is seen in a patient, we look for the first sentence where both (1) FINDING is not empty and (2) the columns NEG and STATUS in FINDING are both empty. This is because all the "normal" findings are expressed either by NON-PATHADJ (*negative, normal*) formatted in the column STATUS or by expressions like *no change, no metastasis, no evidence of metastasis.* If one of the slots in FINDING has an entry other than NEG or STATUS, then it must be an INDICATION, a CHANGE, or a MEDical-FINDING (PART-OF-BODY words will not occur by themselves in the FINDING slot). The format thus standardizes the representation of the important medical information in the sentence, so that this information can be further processed.

RESULTS

To each sentence of our corpus we applied the formatting program, which parsed the sentence, performed certain English transformations on it, and then mapped this structure into the format. This program successfully formatted 176 of the original 188 sentences (94 percent). Table III presents the full format and several examples of formatted sentences.

The full format contains sets of slots for OBSERVE (for doctor+verb: *radiologist noted*), TEST and FINDING. For those sentences that require more than one set of format slots to accommodate their information (e.g., sentences 4 and 5 in Table III), additional sets of format slots are added, each linked to the preceding format by a CONNECTIVE:

FORMAT			CONNECTIVE	FORMAT		
DATA				DATA		
OBSERVE	TEST	FINDING		OBSERVE	TEST	FINDING

In Table III, each row represents a set of format slots; a sentence that requires three sets of format slots (e.g., sentence 5) will therefore occupy three lines of the table.

CONCLUSION

The formatting procedure enables us to convert a natural language corpus into a structured data base. Given a set of x-ray reports in machine readable form, the formatting program maps the input sentences one by one into the tabular format structure. This data base can be used in a variety of ways; we are currently at work on a program to extract various medical statistics from the data base (e.g. number of patients with recurrence of metastasis; time from operation to time of first suspected recurrence of metastasis; location of new metastasis, etc.). It should also be possible to use the data base with a natural language front end to process questions and answer them with information from the data base.

The formatting program is able to convert the natural language material into structured information partly because the material chosen for processing is itself highly structured; however, the formatting relies heavily on a linguistic analysis of each sentence, in order to handle such informationally complex structures as relative clauses, negation, and conjunction.

TABLE III

THE COMPLETED FORMAT: OBSERVE and TEST columns, with examples of formatted sentences.

Labels of the format slots: FORMAT / DATA

NOTE: in the formatted sentences, adjuncts are placed in (); left adjuncts appear above the main word, right adjuncts below it.

	OBSERVE		TEST						
	MD	REPORT	NO-TEST	TEST-VIEW	TESTN	TESTLOC	VERB-DONE	OCCASION	TESTDATE
Findings R sentences, formatted in the columns to the right:									
1. Chest film 6-5 shows enlargement of lesion on right hilum since film of 4-17					film	chest			6-5
2. X-rays taken 3-22-65 reveal no evidence of metastatic disease.					x-rays		taken		3-22-65
3. None this examination			none					(this) examination	
4. Nothing definite is seen that indicates tumor.		(is) seen							
5. The heart, lungs, and bony structures are intact.									

Without a stage of syntactic processing, it would not be possible to determine the scope of negation, the appropriate expansion for conjoined elements, or the referend of the relative pronoun. Because we can process these linguistic structures we can go beyond document retrieval; we are now able to "get inside" the text, to process the actual content.

Our formatting experiment was conducted on a rather simple set of reports which had little paragraph structure and contained specific limited kinds of information. A text that contained several different types of information (requiring several different formats), or had a more complex paragraph structure, with a corresponding increase in intersentential reference, would pose somewhat greater difficulties than the type of material discussed here. Nonetheless there are many instances of natural language material that is both restricted and highly structured (different types of medical reports; weather reports; program specifications in natural language), where this type of procedure would be successful in structuring the information.

Although the specific program described here will process only x-ray reports, the techniques that have been used to obtain the program are general. The string grammar parses English sentences; a few changes enabled it to handle fragmentary note style. The procedure for defining word classes (and selectional restrictions) is based on distributional analysis and can be applied to any language or sublanguage. Part of the procedure for obtaining word classes has been automated in the clustering program;[9] one of our next projects is to complete the automation by adding a program that will convert the parsed sentence into co-occurrence patterns suitable for clustering. The definition of the format was a general technique,

TABLE III—Continued

THE COMPLETED FORMAT: FINDINGS and CONNECTIVE columns.

	FORMAT, continued DATA, continued																					CONNECTIVE		
	FINDING																					NEG-CONN	INDEF-CONN	CONN
	NEG	INDEF-FIND	VERB-ELEMENT		CHANGE-OVER-TIME											STATUS	MED-FINDING	REGION						
			BE-SHOW	INDICA-TION	CHANGE	TIME-PERIOD												POSITION	PART-OF-BODY	STRUC-TURE				
						WHEN	OBSERVE		TEST															
							MD	REPORT	TEST-VIEW	TESTIN	TEST-LOC	VERB-DONE	OCCA-SION	TEST-DATE										
1.			shows		enlarge-ment	since				film				(of) 4-17			(of) lesion	on	(right) hilum					
2.	no		reveal	evi-dence													(of) metastatic disease							
3.																								
4.	nothing (defi-nite)																							rel-clause
	nothing (defi-nite)		indi-cates														tumor							
5.			are													intact			(the) heart					and
			are													intact			(the) lungs					and
			are													intact			bony	struc-tures				

originally developed on a corpus from pharmacology.[4] An overall strategy for mapping parsed sentences into a sublanguage format has been defined, although the transformations themselves are dependent on the target structure (the format) as well as the type of sentence structures in the input. Because each step of our procedure has been based on general linguistic techniques, it should be possible to apply this procedure to convert natural language texts of any sufficiently structured subfield into a structured data base.

REFERENCES

1. Simmons, R., S. Klein and K. McConlogue, "Indexing and Dependency Logic for Answering English Questions," *American Documentation* 15, p. 196, 1964.
2. Harris, Z. S., "Linguistic Transformations for Information Retrieval," *Proc. Int'l. Conf. on Scientific Information* (1958) 2, p. 158, 1959.
3. Sager, N., J. Touger, Z. S. Harris, J. Hamann, and B. Bookchin, "An Application of Syntactic Analysis to Information Retrieval," *String Program Reports* No. 6, Linguistic String Project, New York University, 1970.
4. Sager, N., "Syntactic Formatting of Scientific Information," *Proceedings of the 1972 Fall Joint Computer Conference*, AFIPS Conference Proceedings, Vol. 41, pp. 791-800, AFIPS Press, Montvale, N.J., 1972.
5. Sager, N., "The Sublanguage Technique in Science Information Processing," *Journal of the American Society for Information Science*, Vol. 26, pp. 10-16, 1975.
6. Sager, N., "Syntactic Analysis of Natural Language," *Advances in Computers*, Vol. 8, pp. 153-188, Academic Press, Inc., New York, 1967.
7. Grishman, R., N. Sager, C. Raze, and B. Bookchin, "The Linguistic String Parser," *Proceedings of the 1973 Computer Conference*, pp. 427-434, AFIPS Press, 1973.
8. Anderson, B., I. D. J. Bross and N. Sager, "Grammatical Compression in Notes and Records: Analysis and Computation," paper delivered at the 13th Annual Meeting of the Association of Computational Linguistics, Boston, Nov. 1, 1975, *American Journal of Computational Linguistics*, Vol. 2, No. 4, 1975.
9. Hirschman, L., R. Grishman and N. Sager, "Grammatically-based Automatic Word Class Formation," *Information Processing and Management*, Vol. 11, pp. 39-57, 1975.
10. Hobbs, J. and R. Grishman, "The Automatic Transformational Analysis of English Sentences: An Implementation," to appear in *International Journal of Computer Mathematics*.
11. Sager, N. and R. Grishman, "The Restriction Language for Computer Grammars of Natural Language," *Communications of the ACM*, Vol. 18, pp. 390-400, 1975.