American Journal of Computational Linguistics *Microfiche 36*

PROCEEDINGS.

13TH ANNUAL MEETING

ASSOCIATION FOR COMPUTATIONAL LINGUISTICS

5: MODELING DISCOURSE AND WORLD KNOWLEDGE II. AND TEXT ANALYSIS

Timothy C. Diller, Editor
Sperry-Univac
St. Paul, Minnesota 55101

PREFACE

The fifth and final ACL session was split into two subsessions: one continued the treatment of discourse structure and general knowledge begun in session 4; the other provided a look at several automated text analysis systems. Georgette Silva kindly chaired both subsessions.

Only five of the six talks given are represented in this Proceedings. The paper detailing Salton's talk on automatic indexing was far too extensive to be included on this fiche and hence will be published separately. The paper by Klappholz and Lockman discusses the problems involved in the resolution of cross-sentential reference and sketches an algorithm for their solution. (Note, the closely related paper by Deutsch in Session 4.) Rosenschein addresses the problem of restricting the generation of inferential propositions given a set of beliefs and proposes a structural constraint upon inferencing. Beckles et al. present a man-machine approach to the description of idiolect variations in an environment extraordinarily complex linguistically and sociologically. Brill and Oshika describe a set of programs which permit both batch and interactive processing of orthographic and phonological strings to provide information on frequency, contextual variation, and associational relations. Anderson, Bross, and Sager present a theory of linguistic compression in written texts and describe the results of an implementation of that theory.

Timothy G. Diller, Program Committee Chairman

TABLE OF CONTENTS

SESSION 5A: MODELING DISCOURSE AND WORLD KNOWLEDGE II

Contextual Reference Resolution *David Klappholz and Abe Lockman* 4

How does a System Know When to Stop Inferencing? *Stan Rosenschein* 26

SESSION 5B: TEXT ANALYSIS

Developing a Computer System to Handle Inherently Variable Linguistic Data *D. Beckles, L. Carrington, and G. Warner in collaboration with C. Borely, H. Knight, P. Aquing, and J. Marquez* 40

A Natural Language Processing Package *David Brill and Beatrice T. Oshika* 52

On the Role of Words and Phrases in Automatic Text Analysis *Gerard Salton* (Abstract only) 67

Grammatical Compression in Notes and Records: Analysis and Computation *Barbara B. Anderson, Irwin D. J. Bross, and Naomi Sager* 68

American Journal of Computational Linguistics Microfiche 36 : 68

GRAMMATICAL COMPRESSION IN NOTES AND RECORDS: ANALYSIS AND COMPUTATION

BARBARA B. ANDERSON

Department of Anthropology
University of New Brunswick

IRWIN D. J. BROSS

Roswell Park Memorial Institute
Buffalo, New York

NAOMI SAGER

Linguistic String Project
New York University
2 Washington Square Village, 2B
New York, New York 10012

ABSTRACT

Linguistic mechanisms of compression are used when making notes within a context where the objects and meanings are known. Mechanisms of compression in medical records for a collaborative study of breast cancer are described. The syntactic devices were mainly deletion of words having a special status in the grammar of the whole language and deletion in particular positions of words having a special status in the sublanguage. The deleted forms are described and sublanguage word classes defined. A subcorpus of the medical records was parsed by an existing computer parsing system; a component covering the deletion-forms was added to the grammar. Modifications to the computer grammar are discussed and the parsing results are summarized.

Introduction

All languages have mechanisms of compression. Sentences may be embedded within other sentences by means of nominalization and complementation. Various grammatical transformations involve deletion of certain parts of the sentence.

In medical records, we find entries such as no evidence of metastases, which may be said to be derived from something like There is no evidence of metastases. Such incomplete sentences are not common in the spoken language of the medical records (i.e. dictated reports). However when physicians themselves are required to write material for records, compression mechanisms are commonly used.

Although this paper will deal with a specific corpus, similar devices would often be used for compression in other situations where there is pressure to write as little as possible. Legal, educational, and scientific records where informal notes are kept would be other examples of this class of situations.

The original motivation for this study was to develop effective methods for storing the information in a medical record and to be able to retrieve this information for purposes of research, medical care, or administration. From previous research, the feasibility of verbatim input of dictated narrative has been established. Computerized extraction of the information has been shown to be feasible in a test system ACORN (Automated Coding of Report Narrative). This system has been described in detail in a series of previous papers.[1,2,3]

[1] I.D.J. Bross et al. "Information in Natural Languages: A New Approach". Journal of the American Medical Association, Vol. 207, No. 11, 1969, pp. 2080-2084.

[2] I.D.J. Bross et al. "Feasibility of Automated Information Systems in the User's Natural Language". American Scientist, Vol. 57, No. 2, 1969, pp. 193-205.

[3] P.A. Shapiro and D.F. Stermole. "ACORN (Automated Coding of Report Narrative): An Automated Natural-Language Question-Answering System for Surgical Reports". Computers and Automation, Vol. 20, No. 2, 1971.

For a highly structured medical record where the entries are single words or very restricted sentences, the feasibility of computer-assisted editing and coding has also been established. A procedure for typing in the entries verbatim in a medical record, called 'TICES' (Type-In Coding and Editing System) has been reported elsewhere.[4] However, the third, intermediate class of material cannot be handled by ACORN or by TICES. Therefore, a linguistic analysis of this type of material has been undertaken with the ultimate objective of setting up a comprehensive computer system that can handle almost everything in the medical records.

In the earlier efforts to develop natural language technology, the work was facilitated by the fact that the documents involved were strictly for the transmission of factual information.[5] Such documents are regarded as important both by the persons who are filling them out and by the persons who read them. In this no-nonsense situation where the record may be critically reviewed by the peers of the person who is reporting the information, unambiguous and informative transmission of information is a critical need. Some of the simplicities in the present analysis may be peculiar to this type of situation.

The existence of a subculture with shared training, objectives, and experience may facilitate the note-taking process in somewhat the same way that a person taking notes for himself can somehow be more concise without ambiguity. However, many other note-taking situations would involve a subculture, though not necessarily a medical one, and the findings here might be expected to have some general applicability.

[4] I.D.J. Bross et al. "Unobtrusive Biomedical Data-Input Systems". Bio-Medical Computing, No. 4, 1973, pp. 219-228.

[5] I.D.J. Bross, P.A. Shapiro and B.B. Anderson. "How Information Is Carried in Scientific Sublanguages". Science, Vol. 176, No. 4041, 1972, pp. 1303-1307.

Source of Material

The medical notes discussed here are from the records of the Surgical Adjuvant Breast Project, a nationwide collaborative study involving 36 medical institutions. The records were filled out by medical and paramedical personnel at the participating institutions and centralized at Roswell Park Memorial Institute in a statistical unit under the direction of Dr. Nelson Slack. A sample of approximately 50 was taken from the 2734 case histories of patients in the program and is being used in the linguistic analysis. Each case history ordinarily consists of 3-6 pages of detailed information on the patient's initial status, treatment, pathology report, medical problems, and subsequent fate. When the structured information in the record was excluded, each case history had between 6 and 26 notated items, each item consisting of 1 to 5 partial sentences. While this material is specialized to the purposes of the collaborative study, this type of information is fairly typical of what is found in the usual hospital record.

The notes were typed verbatim using an IBM Mag Card Communicator so as to obtain simultaneously a typed paper document and a record in computer-usable form. This device is used in the data-input system of TICES, an existing system for handling completely structured records. It would presumably be used in any extension of TICES which would handle medical notes. In this analysis the computer was used to reorganize the material in a form more convenient for manual analysis by the linguist.

Anderson analyzed the linguistic structure of the entries in a sample of the medical records involving radiation findings. A discussion of this analysis will take up the next part of the paper. Sager and associates used some of the findings from this study to develop methods for processing these same medical records by computer, adapting a program and grammar which had been developed for parsing science articles. This project will be discussed in the final part of the paper.

Linguistic Characteristics of Medical Notes

Many of the entries on the medical records are in the form of notes which are neither complete sentences nor single word entries, but linguistic strings of an intermediate type, which we will hereafter call fragments. Fragments are a compressed type of linguistic material resulting from various transformations which have the effect of making linguistic strings shorter by reducing or deleting material. The writer of these stretches of material must make his entries brief, in order to save time and effort, but also make them informative and unambiguous. For this reason the deleted material has to be easily recoverable, or in other words it must not contain much information. An analysis of the fragments shows that deletion is mainly of a small class of sentence parts: (1) tense and the verb be (t be); (2) subject, tense and the verb be; (3) the subject; and (4) subject, tense, and verb (V) other than be.

A second characteristic of fragments which makes deleted material recoverable is that both the deleted material and the remainders consist of words in easily defined subclasses, based on both distributional and semantic criteria. These subclasses are easily defined because of the nature of the sublanguage; in general the vocabulary is limited and each word has a limited semantic range. The question on a form which is being answered can also be used as a basis for restoring deleted material.

One of the most commonly deleted items in the medical records is t be (1 and 2). Tense is perhaps the most important information be gives. The deletion of tense in the medical records causes no ambiguity because usually the physician describes the situation at the time of filling out the report. Otherwise he gives the time in a time phrase: x-rays on November 2.

Fragment Types

In Table 1 we list the fragment types, giving an example of each, but not with all occurring word subclasses. The types will first be given according to what material is deleted and then will be further subclassed according to the two highest nodes of the tree structure of the remainder. The material in brackets is the word subclasses which are assumed to have been deleted.

TABLE 1. FRAGMENT TYPES

Material Deleted	Structure of Fragment	Example
1. t be by N-physician	N Ven	no metastatic lesions [were] detected [by physician]
	N Adj	chest films [were] normal
	N P N	patient [was] without cough
	N to V	this form [is] to be used . . .
	N Ving	wound [is] healing well
2. Subject t be	Ven	[N-disease was] aspirated once
	Adj	[N-patient is] dead
	to be Ven	[N-patient is] to be seen by gynecologist
	Ving	[N-patient is] doing well
3. Subject	t V Object	
3a. N-physician Subject		[I] found osteochondritis in rib (5th right)
3b. N-patient Subject		[N-patient] had period one week ago
3c. N-disease Subject (rare)		[N-disease] invades skin [N-disease] seems minor
4. Subject t v		
4a. N-physician t V-discover	Object	[I V-discovered] no bony metastases
4b. N-physician t V-do	Object	[N-physician did] excision of (r) 5th costal cartilage
4c. N-patient t have	Object	[N-patient has] no bone pain

Word Subclasses

The word subclasses should have three characteristics: (1) they should enable deleted material to be recovered, (2) they should make it possible to extract and store informational units such as those in ACORN[6] and (3) they should be defined so that a linguistically unsophisticated person can easily put words into their subclasses.

The word subclasses are based on both semantic and distributional criteria. To a large extent nouns can conveniently be subclassed on a semantic basis and verbs can be subclassed on a distributional basis, according to the subclasses of nouns which they take as subject and object. Due to the nature of the sublanguage there is relatively little overlap (e.g., a given verb is likely to take only one noun subclass as subject) compared to what we would find in the language as a whole.

Two important subclasses of human nouns used in the medical records are N-physician and N-patient. Each has only a few members, but is important because many verbs characteristically take it as subject or object, and also because both, but particularly N-physician, are usually deleted. It is on the basis of the verbs which characteristically take them as subject or object that they can usually be recovered without ambiguity.

Other noun subclasses concern more directly the subject matter of the reports, the concrete objects with which the physician is dealing. Unlike N-physician and N-patient, these classes usually have many members and they are seldom deleted. As with N-physician and N-patient, certain verb subclasses characteristically take them as subject or object.

Table 2 gives some of the word subclasses with examples of each.

[6] Bross et al. "Information in Natural Languages: A New Approach," 1969.

TABLE 2. SOME WORD SUBCLASSES

1.	N-bodypart	abdomen, axilla, bone, breast, cervix, pelvis
2.	N-change	change, elevation, enlargement, gain, increase
3.	N-dimension	pressure, rate, rhythm, size, weight
4.	N-disease	carcinoma, cough, disease, edema, fibrosis
5.	N-exam	biopsy, exam, film, mamogram, scan, x-ray
6.	N-location	area, field, floor, lobe, neck, part, region
7.	N-patient	she, her, patient, lady, woman
8.	N-physician	doctor, he, him, his, I, M.D., radiologist
9.	N-therapy	drug, insulin, medication, medicine, radiation
10.	N-time	date, month, time, visit, winter, year
11.	V-be-equivalent	appear, feel, indicate, remain, represent, seem
12.	V-change	alter, clear, change, enlarge, heal, progress
13.	V-discover	detect, find, identify, note, observe, see
14.	V-patient-object	admit, give, leave, place, readmit, see, transfer, treat
15.	V-patient-subject	complain, come, cooperate, enter, feel, gain, go, have, refuse, show, suffer, take
16.	V-physician-subject	feel, have, place, tell, transfer, treat, see
17.	V-show	show, demonstrate, indicate, reveal, suggest
18.	Adj-bodypart	axillary, bony, clavicular, lumbar, pelvic
19.	Adj-changed	elevated, enlarged, healed, stable, unchanged.
20.	Adj-degree	considerable, extensive, intermittent, little
21.	Adj-discover	absent, evident, known, possible, present
22.	Adj-disease quality	active, bad, benign, degenerative, firm, hard, malignant, metastatic, nodular
23.	Adj-location	adjoining, distal, dorsal, frontal, left
24.	Adj-negative	clear, free, healthy, negative, normal

Computer Parsing of Medical Records[7]

To test the linguistic analysis, a subset of the manually analyzed corpus of medical records was parsed by computer, using the NYU Linguistic String Parser.[8]

[7] I am grateful to Cynthia Insolio and Lynette Hirschman for their help in processing these data. (N.S)

[8] R. Grishman, N. Sager, C. Raze, and B. Bookchin, "The Linguistic String Parser". *Proceedings of the NCC*, AFIPS Press, Montvale, N. J., 1973.

The LSP grammar of English is based on the same linguistic principles as the ACORN grammar. Hence it could also serve to test the feasibility of adding a note-handling capability to the ACORN-TICES system. The LSP sy which was designed for text-processing, was adapted to the parsing of medical records by deleting portions of the grammar which are not required for this type of material and adding a section covering sentence fragments. These changes are described below, followed by the parsing results.

The corpus which was parsed consisted of 12 sections of the Radiation Findings extracted in their order of appearance from the medical records. These sections contained 245 sentences or sentence fragments (word sequences ending in a period). Of these, 37 were complete English sentences and 205 were fragments; 3 were combinations of both types. 21 entries were identical to others in the corpus, accounting in all for 139 of the sentences or sentence fragments. Of the complete sentences, some were quite long, e.g., *Reexamination shows some scarring and thickening over the right apex which is perhaps slightly more evident than it was before, but nothing is seen that is typical of tumor involvement*. Typical shorter sentences are *Chest films on 10-25-68 and 12-14-68 do not show any essential changes since last reports, Liver scan 1-29-69 was normal*. Fragments were, as predicted, of the types listed in Table 1, above, though not all types were represented in the parsed corpus.

Table 3 shows the new definitions or redefinitions which were added to the LSP grammar to cover fragments. These definitions are written in Backus-Naur Form (BNF), as are all the ca. 180 definitions which comprise the context-free part of the LSP English grammar. The BNF definitions are used by the parser to construct a tree representing the structure of the input sentence.

In addition to BNF definitions, the grammar contains restrictions, which test the sentence trees for grammatical and selectional well-formedness.[9] The

[9] For more explanation of the LSP system and grammar, see N. Sager and

TABLE 3. DEFINITIONS ADDED TO THE LSP GRAMMAR
TO COVER SENTENCE FRAGMENTS

```
 1.  <SENTENCE>      ::= <TEXTLET>.
 2.  <TEXTLET>       ::= <OLD-SENTENCE><MORESENT>.
 3.  <OLD-SENTENCE>  ::= <INTRODUCER><CENTER><ENDMARK>.
 4.  <MORESENT>      ::= NULL/<TEXTLET>.
 5.  <INTRODUCER>    ::= NULL.
 6.  <CENTER>        ::= <ASSERTION>/<FRAGMENT>/<IMPERATIVE>.
 7.  <FRAGMENT>      ::= <SA>        (<SOBJBESHOW>/<ASTG><SA>/<NSTG><SA>/
                         <VENPASS>/<NSTG>(<ASSERTION>/<SOBJBESHOW>)).
 8.  <SOBJBESHOW>    ::= <SUBJECT><BE-OR-SHOW><OBJBE><SA>.
 9.  <BE-OR-SHOW>    ::= ↓--↓/NULL.
10.  <ENDMARK>       ::= ↓.↓/↓,↓/↓;↓/↓--↓.
```

starting, or root, definition of the grammar is SENTENCE, so this is the first definition seen in Table 3. In the case of medical records, the unit may be longer than one sentence, but we have retained the root-word SENTENCE and defined SENTENCE in this case to be a TEXTLET (definition 2), which consists of a sentence (called OLD-SENTENCE, definition 3) optionally followed by more sentences (MORESENT, definition 4). The definition of OLD-SENTENCE has the same three elements (INTRODUCER, CENTER, ENDMARK) that the definition of SENTENCE does in the LSP grammar; however, in this case, the INTRODUCER (definition 5) is NULL; the CENTER (definition 6) contains an option FRAGMENT in addition to the options ASSERTION and IMPERATIVE defined in the English grammar (other options of CENTER, e.g. QUESTION, have been deleted); and the ENDMARK (definition 10) contains unconventional punctuation, such as dashes and comma, in addition to the period and semicolon. Since our main interest here is in FRAGMENT (definition 7), we will elaborate on this definition.

R. Grishman, "The Restriction Language for Computer Grammars of Natural Language' *Commun. of the ACM*, 18, 390-400, 1975, and the references cited there.

In defining FRAGMENT, we have used parts of the grammar which were defined independently of the fragment problem. That this is possible is in itself a partial verification of the conclusion from manual analysis that only limited, grammatically specifiable, deletion-forms occur in the fragments seen in notes and records. For example, the dropping of the verb be (type 1 of Table 1) can occur in normal English when a sentence containing the verb be occurs as the object of a verb like find, e.g. We found the chest clear to percussion and auscultation. In the LSP grammar there is an object string defined for such occurrences; it is called SOBJBE (Subject + Object of be). This same string can then be made an option of CENTER to analyze fragments having the same form e.g. Chest clear to percussion and auscultation.

In detail, the definition of FRAGMENT begins with the element SA (Sentence Adjunct). The definition of SA (not shown here) contains 16 options covering all types of sentence modifiers. In this material the most frequent SA is a time expression, usually a date (called PDATE, for optional Preposition + date) or this examination, this visit. Following SA in the definition FRAGMENT are the options proper, naming definitions already in the LSP grammar. The first option SOBJBESHOW (Subject + Obj ect of be or show), corresponds to the second and third structures of type 1 and also occurrences like Chest film no change, which is an expansion of SOBJBE, discussed above. This option covers deletions of the two most common verbs in this material, be and show. The place of be or show (definition 8) in a fragment is either empty or is filled by a dash.

The second and fourth options, ASTG and VENPASS, in FRAGMENT correspond to structures of type 2 in Table 1 (e.g., Negative, felt to be a benign lesion), where the subject, tense and verb be have been dropped. In the LSP grammar, ASTG (Adjective string) is an option of OBJBE, and VENPASS (V-en passive string) is also permitted after be, and in other places. The third option, NSTG (Noun string), is an object of *show*, e.g., *Mild degenerative changes* (from, *X-rays show mild degenerative changes*). It also covers occurrences of the first structure of type 1 (e.g. *No X-rays taken*) where for regularity with more complete entries the passive verb (*taken*) is seen as a right adjunct of the noun. The last option, consisting of NSTG followed by either ASSERTION or SOBJBESHOW, covers such occurrences as *PA and lateral chest 11-5-71 reexamination shows some scarring and thickening over the right apex*. where a noun phrase (*PA and lateral chest 11-5-71*) precedes an assertion about that noun phrase.

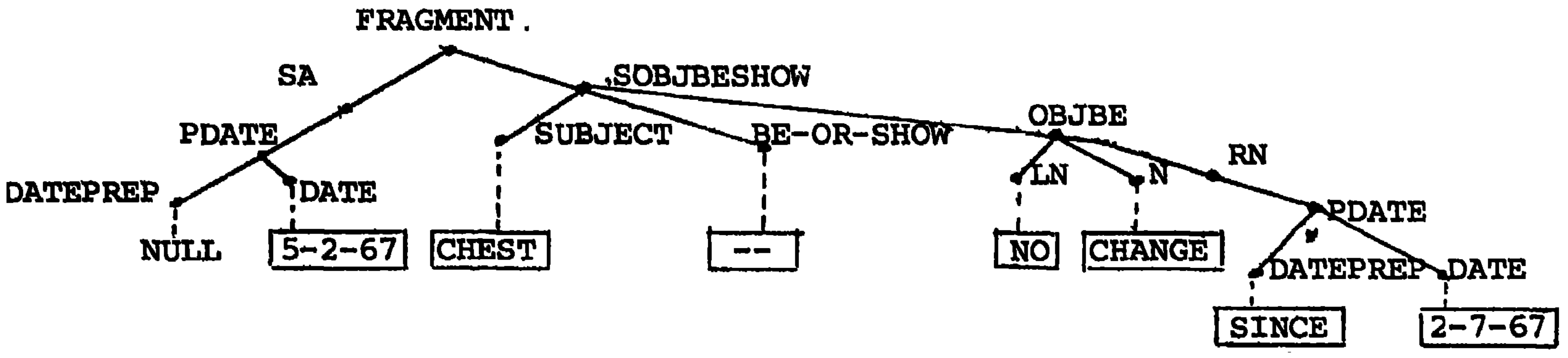
Space permits only a few remarks about these definitions. It was helpful to order the options so that the longer options precede the shorter ones, since some of the shorter options (e.g., NSTG) can have the same form as the first element of the longer ones. This is not required in parsing texts, since in full sentences there is usually no other way of fitting in the remainder of the sentence. Also, in text sentences, many nouns require a preceding determiner, so that compound nouns are not split into separate noun phrases. In this material, determiners are rarely employed, so this constraint cannot be applied. This, combined with verb deletions and the use of commas both in the text and as sentence separators, makes for a great deal of syntactic ambiguity. However, as the next section shows, it was possible to obtain the intended parse as the first output in most cases. This was accomplished without using the subclasses special to the medical material, which are used in a subsequent stage of processing preparatory to information retrieval.

Parsing Results

Parsing output is in the form of a tree, illustrated for a typical fragment in Fig. 1. (Only the nodes mentioned above are shown, plus LN/RN = left/right modifiers of Noun.) The full power of the parser is better illustrated by the long full sentences; but space does not permit presenting them here.

Fig. 1

Parse tree for FRAGMENT = 5-2-67 chest--no change since 2-7-67



A summary of the parsing results is given in Table 4. Of the total 245 sentences, a correct first parse was obtained for 171 or 69.8%, and a first parse adequate for further processing to obtain an "information format" in 213 cases, or 86.9%. The latter statement brings us to the important question of how these parses are to be used.

TABLE 4. PARSING RESULTS

	Number of Sentences	Percentage
Full sentence	37	15.1
Fragment	205	83.7
Full S + Fragment	3	1.2
TOTAL	245	100.0
1st parse correct	171	69.8
1st parse OK for format	213	86.9
2nd or 3rd parse OK for format	14	6.1
No parse or parses 1-3 not OK for format	17	7.0
TOTAL	245	100.0
Average time for 1st parse	5.158 seconds	

The aim in processing natural language notes and records is to arrive at forms for the data which are suitable for computerized information retrieval. The data structures must not change the meaning. This is why syntactic methods are important. Parsing with an English grammar provides the gross structure of input sentences. (The use of English transformations makes the grammatical analysis more refined.) In each specialized technical area, more specific structure is possible, making use of the restricted word usage characteristic of the discourse in the given subject area.[10]

A second stage of processing of this type is now being applied to the parsed corpus of medical records and will be reported in a subsequent paper. A convenient test of the adequacy of the parsing outputs is therefore whether they can serve as input to this second stage of processing (called formatting). It can be seen in Table 4 that a number of "wrong" parses were still adequate as input to the formatting; the segmentation of the sentence into parts was correct even if the parts were assigned an incorrect syntactic status, e.g., object instead of adjunct. Only when the first parse was not adequate for formatting was the sentence rerun to obtain alternative analyses.

The parsing times are a rough indication of the efficiency of the parsing but two points should be kept in mind. (1) The present LSP system is not a production model, but a research tool, with all that implies. (2) A significant fraction of the input sentences were "no data" types, e.g., None this visit. These word sequences were so limited linguistically that a literal formula could serve to recognize them. The experimental use of such a formula cut down parsing times on the no-data entries from about 1.817 to 0.030. However, this formula was not used in the parsing summarized in Table 4.

ACKNOWLEDGEMENT

This investigation was supported by Public Health Service Research Grant number CA-11531 from the National Cancer Institute.

[10] See Ref. 5 and N. Sager, Syntactic Formatting of Scientific Information, Proc. FJCC, AFIPS Press, Montvale, N. J., 1972.