ANNUAL REPORT: AUTOMATIC INFORMATIVE ABSTRACTING AND EXTRACTING

M-21-70-1

March 197

[illegible]

ANNUAL REPORT: AUTOMATIC INFORMATIVE ABSTRACTING AND EXTRACTING

M-21-70-1 March 1970

Annual Progress Report
Office of Naval Research
Contract Nonr 4440(00)

Information Sciences Laboratory
Lockheed Palo Alto Research Laboratory
LOCKHEED MISSILES & SPACE COMPANY
A Group Division of Lockheed Aircraft Corporation
Palo Alto, California 94304

PRÉCIS

RESEARCH PROGRESS REPORT

Title: "Annual Report: Automatic Informative Abstracting and Extracting," Annual Progress Report, Office of Naval Research, Contract Nonr 4440(00).

Authors: L. L. Earl and H. R. Robison

Background: This investigation is concerned with the development of automatic indexing, abstracting, and extracting systems. Basic investigations in English morphology, phonetics, and syntax have been pursued as necessary means to this end. Experimental indexing and extracting systems are now being developed.

Condensed Report Contents: Part I of this report documents several experiments in automatic extracting and one experiment in automatic indexing. Nine chapters, each from a different technical book, were used as the text corpus for all the experiments. In the first experiment, an attempt was made to construct a sentence dictionary of syntactic sentence types, for distinguishing extract-worthy sentences, but it proved unrewarding. Nevertheless, the results indicated that sentence typing might be used in a screening process in conjunction with other extracting techniques. The later attempts to combine syntactic and statistical criteria in the choice of extract sentences and index phrases proved more rewarding. The sentences selected by the extracting algorithm were representative and are presented for the reader to peruse. The noun phrases selected by the indexing algorithm compared favorably with the back-of-the-book index phrases. There is every indication that satisfactory back-of-the-book indexes could be produced automatically, with post-editing to delete superfluous items.

Part II reports on the relationship between English word government and the problem of multiple meaning in natural-language processing. A set of English words is discussed each of which has the ability to convey its various semantic meanings by the use of certain common syntactic units such as prepositions. Because these common syntactic units occur adjacent to – though not necessarily contiguous with – the word whose meaning they specify, the entire complex (here called a semantic structure) is computer-recognizable; thus the phenomenon described opens up the possibility of teaching the computer to make semantic distinctions. The various syntactic units which are used to make semantic distinctions are discussed separately and examples are given. Because prepositions play an important role in making semantic distinctions, a section on prepositional semantics is included.

For Further Information: The complete report is available in the major Navy technical libraries and can be obtained from the Defense Documentation Center. A few copies are available for distribution by the authors.

LOCKHEED PALO ALTO RESEARCH LABORATORY
LOCKHEED MISSILES & SPACE COMPANY
A GROUP DIVISION OF LOCKHEED AIRCRAFT CORPORATION

FOREWORD

This report marks the completion of the sixth year in which the Office of Naval Research has contributed support to research in the Information Sciences at the Lockheed Palo Alto Research Laboratory of the Lockheed Missiles & Space Company. During the first year of the program, a major part of the effort went into establishment of a word-data base. The English Word Speculum, which has been distributed to ONR program participants, illustrates the nature of this data base. In the second and third years, this data base was exploited in the development of a computer program for the automatic assignment of parts of speech to English words. Also during these years, it was demonstrated how an English/Russian phrase data base can be used to develop a technique for obtaining English indexes from untranslated Russian text. In the third and fourth years, a new data base of sentences with assigned parts of speech was created for investigation of the abstracting and extracting process. Also during the third and fourth years, experiments in the compilation of a "sentence dictionary" of syntactic types began, and compilation of English syntactic word government tables began. These activities were continued in the fifth year, along with development of a parsing program, the initiation of some extracting experiments on some technical text, and an experiment in automatic indexing of a medical book. This year the "sentence dictionary" experiment was concluded, the extracting experiment was completed, and a frequency-syntax method of automatic indexing was conceived and tested. Also during this year the concept of English syntactic word government was expanded, and compilation of the tables continued.

Part I of this report is concerned with the sentence dictionary and indexing and extracting experiments. Part II is concerned with the concepts of English word government.

The group at Lockheed takes this oppotunity to express its thanks for the continuing support and encouragement given by the Information Sciences Branch of the Office of Naval Research.

ILLUSTRATIONS

CONTENTS

Part I

EXPERIMENTS IN THE USE OF SYNIACTIC INFORMATION IN AUTOMATIC EXTRACTING AND INDEXING

LOCKHEED PALO ALTO RESEARCH LABORATORY
LOCKHEED MISSILES & SPACE COMPANY
A GROUP DIVISION OF LOCKHEED AIRCRAFT CORPORATION

Section 1
BACKGROUND AND THEORY

1.1 PHRASE STRUCTURE IN "SENTENCE DICTIONARY" EXTRACTING

It would be very convenient for automatic extracting if a relationship existed between the syntactic structure of a sentence and its significance or usefulness in forming an extract. This would be particularly true if a "sentence dictionary," a dictionary of syntactic sentence types, could be compiled and then a separation of significant from nonsignificant sentences could be accomplished on the basis of these sentence types. To test such a possibility, an experiment was devised for building a trial sentence dictionary using part-of-speech strings to characterize a sentence syntactically, and a previously formed index to establish whether a sentence was to be considered significant or not. This experiment is described and documented in three previous annual reports (Refs. 1, 2, 3). The results will be summarized in the following paragraphs as background for the continuing sentence dictionary experiment.

According to the sentence dictionary hypothesis, a large group of sentences, as representative of the language as possible, are to be processed, classified as "indexible" or "nonindexible," and assigned a syntactic structure. The hope is that when these structures are sorted, or ordered, it will be found that like structures have like index classifications. If this is so, and if there are not too many different structures possible in the language, the structures can be ordered into a "dictionary" of sentence types, each classified as indexible or nonindexible, so that when other sentences are assigned a sentence structure, a dictionary look-up can be used to classify them as indexible or nonindexible. (The dictionary really need contain sentence types of only one classification, either all indexible or all nonindexible sentences.)

Testing ɔf the validity of the sentence dictionary hypothesis boils down to answering two questions:

- Is there a system of syntactic classification which will separate indexible from nonindexible sentences?
- Does such a system produce a manageable number of sentence types?

In the compilation of a trial sentence dictionary, several derivative and interrelated questions had to be faced immediately.

- What sentences can be used as a data base in constructing a sentence dictionary?
- How should sentences be identified as indexible or nonindexible?
- What method of syntactic classification should be tried?
- How many sentences must be used in the data base in forming the sentence dictionary so that all, or nearly all, English syntactic types will be represented?

The initial answers to these questions need not be the final answers. The experiment, in fact, consists of working to arrive at optimum practical answers to these questions so that the validity of the sentence dictionary hypothesis can be evaluated. The following paragraphs give the answers adopted for the initial experiment.

- <u>What sentences can be used as a data base in constructing a sentence dictionary</u>? The sentence dictionary hypothesis depends on the further hypothesis that well-formed English sentences in scientific exposition follow a finite number of syntactic patterns, which will occur repeatedly in text, so that if enough text is examined, the patterns will emerge. Thus the question concerning the number of sentences to be chosen becomes more important than the sentence source, although it seems obvious that the more diverse the sample, the more economically the objective can be met. To choose text for the initial experiment, books were picked at random from the Palo Alto Library of the LMSC Technical Information Center, and one chapter was used from each book. It was hoped that different books would reflect different styles, and that using the running text from one chapter would give a representative sample of the construction used by that author.

- How are sentences to be identified as indexible or nonindexible? Since books are usually published with an index, and these indexes have long been accepted by the public, it seemed simplest to begin by using these indexes in distinguishing which sentences were to be regarded as significant and, therefore, which sentence types were to go into the sentence dictionary as indexible. Sentences were designated indexible if they contained a phrase listed in the index, and nonindexible if they di not. The difficulties encountered in making such a designation are discussed on pages 1 – 5 and 1 – 6 of Ref. 1.

- How are sentences to be syntactically classified? It seemed reasonable to begin b classifying sentences into as fine a classification as possible, because the finer the syntactic classification and the larger the number of syntactic sentence types, the greater the chance that a partition into indexible and nonindexible types can be made. Therefore, the sentences were initially characterized as a string of part-of-speech possibilities. Because dictionary look-up of words is costly, a study of the graphemic form of words was made, and an algorithm was devised to assign part-of-speech codes to words automatically (Ref. 4). Figure 1 in Ref. 2 shows sentences with their assigned part-of-speech strings.

- How many sentences must be used in forming the sentence dictionary? Unfortunately, this question and the third question, how the sentences are to be syntactically classified, are interrelated. The number of sentences which must be processed in order to encounter nearly all syntactic types depends on the system of syntactic classification used, and the practicality of a syntactic classification system depends on the number of sentence types produced by the classification system. As with the other questions, the answer is really to a large extent the object of the experiment, but always there must be a starting place. Nine chapters from nine texts, listed in Ref. 1, were used in the initial experiment; this sample of 3,216 sentences was thought sufficient to indicate the further directions in which the experiment should proceed.

The results of the initial sentence dictionary experiment are detailed in Section 2 of Ref. 2. From these results it became clear that representing a sentence by part-of-speech strings makes too fine a distinction between sentences. Nearly all the sentences of the data base (3064 out of 3216) had a unique part-of-speech pattern. Obviously, another method of classifying sentences had to be chosen to decrease the total number of sentence types and to increase the probability of duplicate types. It was decided to try classifying the sentences according to their phrase structure, with the hope that there would be a significant number of sentences of like phrase structure. To do this, it was necessary to develop an algorithm for automatic syntactic analysis, based on a phrase-structure grammar. The syntactic analysis algorithm developed was based on the needs of this sentence dictionary application and of two other applications, as described in subsections 1.2 and 1.3 of Ref. 3. In all three applications, identification of noun and verb phrases was of primary importance; the need for identification of higher level structures (such as participial and prepositional phrases and clauses) and of modification patterns had not yet been established. Since the identification of noun and verb phrases and infinitives is basic, is a large part of the parsing process, and is necessary for higher level analysis, it was decided to limit the first parsing efforts to identification of these elements and to resolution of the role of participles in the sentence.

Thus, the second experiment in syntactic extracting was to be exactly analogous to the first, except that the syntactic classification was now in terms of both phrases and parts of speech rather than in terms of parts of speech alone. The structure of each sentence was defined by the parsing program, BPHRAS, in terms of its noun phrases, verb phrases, infinitives and gerunds, with the connecting function words represented by their parts-of-speech strings, as shown in Fig. 1. The parsing program, BPHRAS, described and documented in Refs. 2 and 3, accomplishes this by grouping the words of the sentence into possible units and then resolving the conflicts among units until each word is assigned to but one unit and the sentence is represented by a single string of elements. This means that a choice must be made among alternative structures, a choice that is sometimes in error. The usual ambiguities are between noun and verb phrases, between infinitives and prepositional

Accordingly	→	adjective-adverb
,	→	punctuation
with	→	preposition
Jefferson	→	noun phrase
,	→	punctuation
they	→	noun phrase
feel	→	verb phrase
that	→	noun phrase (conjunction)
the maintenance	→	noun phrase
of	→	preposition
a competitive free enterprise system	→	noun phrase
is	→	verb phrase
a basic requirement	→	noun phrase
for	→	preposition
continuing	→	present participle
a democratic government	→	noun phrase

Fig. 1 Sample Parsing Results

phrases, and among different usages of participles. An effort was made to achieve 97% accuracy, i.e., to identify correctly 97% of the total number of elements (including function words). A small sample of about 50 sentences was worked over until this goal was achieved on that sample, and, without further modification, on about 125 more sentences from the same text. This does not mean that accuracy will be that high with general text, particularly since some errors are due to inaccuracies or inadequacies in the part-of-speech program. Many of these were eliminated for the text in question by adding such words in that text to a dictionary of words to be corrected.

With the parsing program ready to use, two other programs were prepared for the second sentence dictionary experiment, one for sorting the new sentence structure (phrase-structure SORT), and one for comparing the sorted structures to select sentences with duplicate structures (DUPPHRAS). Development of these programs was discussed in section 2.1 and 2.2 of Ref. 3. Section 3.2 of Ref. 3 documents the phrase-structure SORT, but DUPPHRAS was not completed until this year and is documented in section 3.2 of this report. The second sentence dictionary experiment was completed this year and the results are discussed in section 2.1 of this report.

1.2 PHRASE STRUCTURE IN FREQUENCY-SYNTAX INDEXING AND EXTRACTING

In the early years of this contract, the first algorithm combining syntactic and frequency criteria was developed for automatically selecting index items. This is described briefly in the March 1966 Annual Report (Ref. 5). In these experiments, the data corpus was reduced by selecting only certain structural items (subjects, objects of verbs and infinitives, with genitive modifiers); then this reduced corpus was subjected to statistical counts. An algorithm was developed, which, on the basis of these counts, chose certain words for a first-level index. Then all larger structural items containing these words were included in a second-level index. This algorithm was tested on five text excerpts, but because of deficiencies in the parsing program (loopholes in parsing logic led to trouble every 50 sentences or so), it was impossibie to pursue the experiment. Nevertheless, the results on the five excerpts promising, and proved a useful guide in the development of present algorithms for both indexing and extracting.

In reviewing the early work, it was first of all obvious that a new syntactic analyzer was needed. Others in current use were also unsuitable, being either too limited, too prone to trouble, too time-consuming or too indefinite in the assigning of sentence structure. Concurrently developing needs in the sentence dictionary experiment led to the development of a much simpler parsing program suitable to both purposes, as described in section 1.1. It was also clear in reviewing the early work that it is indeed desirable to reduce the text by syntactic criteria, both to eliminate the high-frequency function words, which would otherwise need to be eliminated by a more complicated statistical analysis, and simply to econo. ize on the frequency-counting necessary. Instead of extracting higher level elements, as in the earlier experiments, it seemed simpler and very possibly more advantageous to begin by selecting all noun phrases, excluding adverb and function words within the noun phrase. This is a good place to start because it is more inclusive; later experiments can cut down to higher level elements if that is desirable. To illustrate, all the nonfunction words of all noun phrases of several sentences have been underlined.

Note that besides reducing the text, this will reduce the possibility of introducing adverbs, adjectives, and verbs which could be confused with significant nouns. Luhn (Ref. 6), for instance, counted all words with the same first six letters as the same, and worried somewhat about confusing such words as "differ, different, or differently" with a technical word like "differential." In this test, "differ" and "differently" would never appear in the reduced text.

Frequency counts have been used both in extracting and in indexing techniques, and it was decided to develop algorithms for both, based on the reduced text. The extracting algorithm was developed first. The evolution of this algorithm is described in section 2.3 of last years annual report (Ref. 3). The algorithm itself is summarized in the following paragraphs. Testing of the extracting algorithm is described in section 2.2 of this report, and development and testing of the indexing algorithm is described in section 2.3 of this report.

The extracting algorithm developed required seven steps, as follows:

(1) All the nonfunction words in noun phrases containing more than one word were extracted from the text, with the text, page, and sentence number from which they came.

(2) This list was alphabetized, and a frequency count made of the words.

(3) The frequency counts were then ordered (descending order) and the difference from one count to the next was calculated and stored.

(4) The percent of each frequency count of the total of all frequency counts was calculated and stored with the cumulative percent to that point.

(5) Four points in the frequency count table were then calculated and the four frequencies were ordered in ascending order. Of these, one was chosen as the cutoff frequency. The second was taken if that point had a cumulative percent less than an arbitrary value V2. If not, then the third point was taken if it had a cumulative percent less than V2. If not, then the fourth point is taken. (Calculation of the four points will be discussed later.)

(6) All frequencies lower than the cutoff frequency were discarded. The words making up the remaining frequency counts were extracted from the original list and ordered (ascending) according to their associated page and sentence number.

(7) The word-sentence number combinations in this list were examined, and sentences were located and printed from the text for all sentences in which:

(a) there were three or more high-frequency words

(b) there were two high-frequency words, if that pair had not already occurred together in a sentence

Some of these steps need some elaboration. In step (5), four points within the frequency count table are identified. Point 1 is the point at which the cumulative percent [calculated in step (4)] is less than or equal to an arbitrary value V1. Point 2 is the point at which the largest differences in frequencies occur. Point 3 is the point at which the largest difference below point 2 occurs. Point 4 is that frequency equal to or just larger than the largest frequency divided by 2. Points 2 and 3 represent points at which there is a steep slope in the frequency curve. Points 1 and 4 represent points at which an arbitrary cutoff might be made.

Step (7) is necessary to reduce the volume of sentences selected. Taking sentences with three or more high-frequency words is similar to taking the highest weighted sentences in schemes like Luhn's. Taking sentences with two high-frequency words only if it is the first occurrence of that word pair is an attempt to pick up each new idea, yet avoid redundancy in the sentences selected.

Section 2
1969 – 1970 PROGRESS

2.1 THE "SENTENCE DICTIONARY" EXPERIMENT

By the end of the 1968 – 69 year, the parsing program BPHRAS, the phrase-structure SORT, and the compare program DUPPHRAS were complete except for the checkout of DUPPHRAS. (See section 1.1 for theory and background information.) This checkout was completed, and then the second sentence dictionary experiment was carried out, in steps analogous to those of the first:

(1) The processed sentence files with part-of-speech assignments were reprocessed to form a file with the part-of-speech information replaced by phrase structure information (BPHRAS program).

(2) The sentence file (or files) was ordered on the phrase structure information (phrase-structure SORT).

(3) The ordered files were examined by computer to locate and print sentences with the same phrase structure (DUPPHRAS program).

In this experiment, nine texts were processed by BPHRAS, and these were then sorted together in two groups, one with five texts and one with all nine. (Reference 7 lists the nine books used; one chapter was used from each.) Again we were interested in testing the validity of the sentence dictionary hypothesis by answering the questions:

- Is there a system of syntactic classification which will separate indexible from nonindexible sentences?
- Does such a system produce a manageable number of sentence types?

Therefore, again we were interested in the number of duplicated structures, the extent to which they are associated with the same significance code, and any significant trend in the type of sentence being printed out. Table 1 summarizes the same statistics for phrase patterns which was given in Ref. 2 for part-of-speech patterns. Table 2 compares the results of all nine chapters for the part-of-speech and phrase patterns.

Table 1

STATISTICS OF PHRASE PATTERNS IN TEXT

Item	Number of Chapters in Data Base	
	5	9
(1) Number of total patterns represented by more than one sentence	22	35
(2) Same as (1), with a consistent code	11	15
(3) Number of duplicate patterns common to more than one article	17	26
(4) Same as (3), with a consistent code	7	11
(5) Number of one-of-a-kind patterns	1840	3026
(6) Number of total unique patterns	1866	3061
(7) Ratio of the number of one-of-a-kind to number of total unique patterns	0.982	0.988

Table 2

COMPARISON OF PART-OF-SPEECH AND PHRASE PATTERNS

Item	Part-of-Speech Patterns	Phrase Patterns
(1) Number of total patterns represented by more than one sentence	34	35
(2) Same as (1), with a consistent code	23	15
(3) Number of duplicate patterns common to more than one article	12	26
(4) Same as (3), with a consistent code	5	11
(5) Number of one-of-a-kind patterns	3064	3026
(6) Number of total unique patterns	3098	3061
(7) Ratio of the number of one-of-a-kind to number of total unique patterns	0.992	0.988

These are disappointing results. Both levels of syntactic classification have failed to separate indexible from nonindexible sentences. The first and most elementary classification provided far too many sentence types, so a higher level of classification – into phrase patterns – was tried. This was expected to merge some types, reducing the total number. However, examination of Table 2 shows that even with the phrase pattern classification there are still far too many unique patterns, and furthermore that the consistency of index codes is dropping with the number of unique patterns.

Close examination of Table 2 shows that there are not significantly more duplicate patterns when using phrase patterns than with part-of-speech patterns, although there are a few more sentences representing each pattern. (The number of duplicate patterns common to more than one article is up by more than a factor of 2, although in numbers it is only 14 more patterns.) At the same time, however, the consistency of index codes has dropped from 68% to 43% for all patterns and from 52% to 42% for patterns common to more than one article. Note that the ratio of the number of one-of-a-kind to the number of total unique patterns has dropped by only 0.004. It seems fair to say that indexible and nonindexible sentences cannot be distinguished by structure alone.

All that remains is to try to evaluate the experiment from viewpoints other than that of a sentence dictionary. There are some general observations to be made. One is that over half (109 of 182 or about 60%) of the duplicated patterns are titles, and that the rest of the duplicated patterns are short and rather trivial sentences. It is this mixture of titles and trivial sentences which accounts for the inconsistency of the index codes. There are only 13 sentences in the duplicated pattern group which have an I index code and are not titles.

Titles have been regarded as good extract material; however, it may be that they are really good index material and rather poor extract material. Compare, for example, the title "Controlling Inflation and Ensuring Stability" with the sentence, "...Public concern about inflation enters into policy considerations about competition in two ways: first is the influence of competition on inflation and on the competitive drive for efficiency...." Compare also the title "Surface Processes in Liquid Polymer Formation" with "...The various biochemical theories of liquification fall conveniently into three categories: (a) liquid arises in the cell wall by direct transformation of other wall components, (b) liquid arises from precursors which diffuse centripetally from their point of origin in the cambium and become incorporated into the walls of xylim and phloem, (c) liquid arises from cytoplasmic precursors formed in differentiating cells and is subsequently incorporated into the wall...." In both cases, the titles seem better and much more concise indexes to the subject matter; the sentences seem to give a better clue to the style and point of view of the author while continuing to convey the subject matter.

These observations point the way to two possible experiments. For one, an attempt could be made to use function word counts or frequency counts to differentiate between the titles and the trivial sentences whose syntactic forms turned up together in the phrase structure experiments. This would allow titles to be picked up without requiring any keypunched flag on titles, and would also allow automatic rejection of the less meaningful titles such as "Results," if that were desirable. For another, experiments might be conducted in choosing sentences (or perhaps simply in preferring sentences chosen by other criteria) which have the most complicated or the longest phrase structure, on the grounds that if the simplest sentences (other than titles) are trivial, the longest will encompass the most significant ideas, or will be acting in a summarizing capacity.

2.2 FREQUENCY-SYNTAX EXTRACTING EXPERIMENT

After the frequency-syntax algorithm was developed, as described in Ref. 3 and summarized in section 1.2 of this report, it was tested on a chapter from each of four texts, 1, 3, 7, and 9 of Ref. 7. Text 7 was used in the development of the algorithm and so cannot strictly speaking be called a test case.

In Step (5) of the algorithm (see section 1.2), four points within the frequency count table are identified. Points 1 and 4 represent points at which an arbitrary cutoff might be made, the first the point at which the cumulative percent reaches an arbitrary value V1, the fourth the point at which the frequency count reaches the largest frequency divided by 2. Points 2 and 3 represent points at which there is a steep slope in the frequency curve. The tests of the algorithm were each run five ways; each of the four points was taken as a cutoff point, and a point was automatically chosen from among them by the algorithm and taken as the cutoff point.

The results of this algorithm on the four texts can be evaluated in many ways. Table 3 gives some interesting statistics for the tests in which the cutoff point was chosen automatically. The text number is the number of the text in Ref. 7. The selection percent is the ratio of sentences in the extract to those in the text. The coselection percent is the ratio of the number of selected sentences designated as indexible to the total number of selected sentences. The acceptable selection percent is the ratio of the number of sentences deemed acceptable (by the author) in the extract to the total number of selected sentences. Sentences were deemed unacceptable if they seemed trivial in content, if they contained unclear antecedents, or if they were redundant with another better sentence or one marked indexible.

Table 3

SYNTACTIC-STATISTICAL EXTRACTING RESULTS

Text No.	Chapter Name	Number Sentences Extracted	Number Sentences Text	Selection %	Co-Selection %	Acceptable Selection %
1	Mao's Strategy	31	248	12.5	48.4	87.1
3	Art, Life & Experiment	24	414	5.8	58.3	70.9
7	Basic Public Goals	18	479	3.8	66.7	95.0
9	Cell Wall Dynamics	15	341	4.4	40.0	86.6

In one of the texts, number 3, failure to identify the words "its" and "no" as function words caused these words to occur among the high-frequency words. Fortunately, because of the use of unique pairs or three or more high-frequency words [step (7)], the results were not degraded unduly, although the acceptable selection percent is lowest for this test. "Its" also occurred among the high-frequency words in text 1 and may be partly responsible for the high selection percent in that case. The co-selection percent and acceptable selection percent are highest for text number 7 becau it was used in the development of the algorithm.

The high-frequency words selected may be of interest and are given in Fig. 2 for eac of the four texts. The extract sentences for two of the texts are given in Fig. 3. The sentences marked by * are indexible; that is, they contain words found in the back-of-the-book index. In the first 15 sentences of each text, the high-frequency words within the sentences are underlined. In the "Cell Wall Dynamics" text, nearly all the sentences were chosen because they contained three or more high frequency words. Note that the repetition of a high frequency word was counted as another instance of a high frequency word. In the "Roots of Mao's Strategy" text, 9 of the firs 15 sentences also were chosen because they contained 3 or more high frequency words though with 5 rather than 3 high frequency words retained, there are 15 rather than 6 possible word pairs.

The extract results shown in Table 3 and Fig. 3 are mildly encouraging. All four extracts cover the territory of the text rather well and convey the subject matter and tone of the text. For screening purposes, it is felt that the extracts produced by this algorithm are quite acceptable.

Table 3 shows the extract statistics when the frequency cutoff point was chosen automatically from among four points, each of which was also chosen automatically. Tests were also completed in which points 1 through 4 were arbitrarily used as the frequency cutoff point in step (5) of the sentence selection algorithm. The choice of different points varies the number of high-frequency words selected in step (5) for use in sentences in step (7), thus varying the selection, coselection, and acceptable

Text	Chapter Name	Word	Frequency
1	Mao's Strategy	Chinese	74
		communist	58
		its	50
		world	45
		communists	44
3	Art, Life & Experiment	heart	35
		its	32
		blood	30
		science	27
		no	25
		body	23
		history	19
7	Basic Public Goals	competitive	52
		economic	42
		power	36
		policy	33
		public	29
		prices	28
9	Cell Wall Dynamics	acid	86
		cell	64
		wall	36

Fig. 2 Selected High-Frequency Words

- IT IS NOT EASY TO GRASP THE UNDERLYING MOTIVATIONS AND OUTLOOK OF THE LEADERS OF ANY MAJOR NATION, AND IN SEEKING TO DEFINE THE PURPOSES OF THE CHINESE COMMUNISTS THE DIFFICULTIES ARE COMPOUNDED BY MANY CULTURAL, IDEOLOGICAL, AND HISTORICAL BARRIERS.

  BOTH AS ASIAN NATIONALISTS AND AS COMMUNIST REVOLUTIONARIES, MAO AND HIS FOLLOWERS VIEW THE WORLD IN TERMS WHICH ARE LITTLE UNDERSTOOD IN THE WEST.

  AND THE OFFICIAL CHINESE COMMUNIST PRESS HAS STATED BLUNTLY THAT NO SOLUTION OF ANY INTERNATIONAL PROBLEM, ANY ASIAN PROBLEM IN PARTICULAR, IS POSSIBLE WITHOUT THE PARTICIPATION OF THE CHINESE PEOPLE'S REPUBLIC.

  THE HISTORY OF CHINA, THE TRADITIONAL CHINESE VIEW OF THE WORLD, AND THE FORCE OF MODERN NATIONALISM ALL PLAY A PART IN MAKING THIS A DRIVING FORCE IN PEKING'S POLICY.

- UNTIL THE MODERN PERIOD THE CHINESE REGARDED THEIR COUNTRY AS THE CENTRAL KINGDOM, THE CENTER OF THE CIVILIZED WORLD, SURROUNDED BY STATES WHICH EITHER ACCEPTED A SUBORDINATE TRIBUTARY RELATIONSHIP OR WERE CONSIDERED INFERIOR NATIONS OUTSIDE THE PALE OF THE CHINA-CENTERED CIVILIZED WORLD.

  ALTHOUGH THE OLD BASIS FOR THIS WORLD VIEW HAS FADED, MOST CHINESE, INCLUDING THE COMMUNISTS, STILL HARBOR A FEELING OF SUPERIORITY OVER THEIR NEIGHBORS (A FEELING NOT DISSIMILAR TO THAT WHICH MANY WESTERNERS HAVE FELT TOWARD THE NON-WESTERN WORLD IN THE MODERN PERIOD.)

- THE COMMUNIST PARTY'S VICTORY WITHIN CHINA WAS DUE IN NO SMALL MEASURE TO ITS SUCCESS IN APPEALING TO NATIONALIST SENTIMENT BOTH DURING THE WAR WITH JAPAN AND AFTERWARD.

- ALTHOUGH NATIONALISM IS ONE OF THE DRIVING FORCES BEHIND THE CHINESE COMMUNIST ACTIONS, IT IS CLEARLY THE IDEOLOGICAL CONVICTIONS OF PEKING'S LEADERS WHICH SHAPE THEIR PRESENT VIEW OF THE WORLD, MOLD THEIR STRATEGY, AND PROVIDE THE RATIONALE FOR BOTH THE ENDS AND MEANS OF THEIR POLICY.

- THEY ACCEPT LENINISM (LENIN FOR PRESENT-DAY COMMUNISTS IS A MORE IMPORTANT PROPHET THAN MARX) AND ITS KEY DOCTRINES: THE IDEA THAT IMPERIALISM IS THE FINAL STAGE OF CAPITALIST DECAY; THE STRESS UPON THE IMPORTANCE OF COLONIAL AND SEMICOLONIAL COUNTRIES IN THE REVOLUTIONARY STRUGGLE AGAINST CAPITALIST COUNTRIES; THE IMPORTANCE OF ANTI-IMPERIALISM IN WORLD REVOLUTION; THE IDENTIFICATION OF THE PROLETARIAT WITH A DISCIPLINED, ELITE, DEMOCRATIC-CENTRALIST PARTY WHICH MUST LEAD THE REVOLUTION THE RECOGNITION OF THE POTENTIAL REVOLUTIONARY IMPORTANCE OF THE PEASANTRY; AND THE IDEA THAT IN LESS DEVELOPED COUNTRIES REVOLUTION MAY GO THROUGH A BOURGEOIS-DEMOCRATIC PHASE.

  THE CHINESE COMMUNISTS ALSO ACCEPT MOST OF THE STALINISM AND ITS DOCTRINE ON THE TOTALITARIAN ORGANIZATION OF STATE POWER, ON STATE-DIRECTED ECONOMIC DEVELOPMENT, AND ON OTHER QUESTIONS SUCH AS THE PROBLEM OF DEALING WITH NATIONAL MINORITIES.

- IN APPLYING MARXIST-LENINIST DOCTRINES TO THE CHINESE SCENE, THE COMMUNISTS IN CHINA, AND ABOVE ALL MAO TSE-TUNG, HAVE DEVELOPED IDEOLOGICAL FORMULAS WHICH MAY LEGITIMATELY BE CALLED MAOISM, ALTHOUGH THE CHINESE THEMSELVES REFER TO THEM MERELY AS THE THOUGHT OF MAO-TSE-TUNG.

  IT IS NOT POSSIBLE HERE TO DISCUSS ALL THESE DOCTRINAL CONCEPTS, OR TO TRACE IN DETAIL THE PROCESS OF INTERPRETATION AND ADAPTION TO WHICH THEY HAVE BEEN SUBJECTED IN COMMUNIST CHINA AND ELSEWHERE WITHIN THE COMMUNIST BLOC.

  THE CONCEPT OF THE UNITED FRONT, WHICH PLAYED AN IMPORTANT PART IN THE COMMUNIST VICTORY WITHIN CHINA, IS OFTEN APPLIED BY PEKING IN ITS STRATEGY ABROAD, BOTH IN ITS RELATIONS WITH OTHER GOVERNMENTS, AND IN ITS RELATIONS WITH OTHER GOVERNMENTS, AND IN ITS NONOFFICIAL REVOLUTIONARY CAMPAIGNS TO MOBILIZE PRESENT OR POTENTIAL FOLLOWERS.

  TO THIS POLE COMMUNIST CHINA IS ATTACHED BY WHAT THE COMMUNISTS LABEL AN INDESTRUCTIBLE FRIENDSHIP, AND, ALTHOUGH WITHIN THE COMMUNIST BLOC PEKING HAS RISEN TO A POSITION OF ASSOCIATE LEADERSHIP WITH MOSCOW, IT STILL ACKNOWLEDGES THE PRIMACY OF THE SOVIET UNION.

  HOWEVER TOPSY-TURVY THIS VIEW OF THE WORLD MAY SEEM TO PEOPLE OUTSIDE THE COMMUNIST BLOCK, IT IS RIGIDLY UPHELD WITHIN THE COMMUNIST ORBIT.

Fig. 3a Extract Sentences: "Mao's Strategy"

- LIU SHAO-CH'I HAD WRITTEN IN 1948 THE WORLD TODAY HAS BEEN DIVIDED INTO TWO MUTUALLY ANTAGONISTIC CAMPS. ON THE ONE HAND, THE WORLD IMPERIALIST CAMP, COMPOSED OF AMERICAN IMPERIALISTS AND THEIR ACCOMPLICES, THE REACTIONARIES OF ALL COUNTRIES OF THE WORLD. ON THE OTHER HAND, THE WORLD ANTI-IMPERIALIST CAMP, COMPOSED OF THE SOVIET UNION AND THE NEW DEMOCRACIES OF EASTERN EUROPE, AND THE NATIONAL LIBERATION MOVEMENTS IN CHINA, SOUTHEAST ASIA AND GREECE, PLUS THE PEOPLE'S DEMOCRATIC FORCES OF ALL COUNTRIES OF THE WORLD.

  THE CHINESE COMMUNISTS FEEL THAT, AS ASIANS, THEY CAN AND SHOULD PLAY AN ESPECIALLY IMPORTANT ROLE IN ALIGNING ANTICOLONIALIST AND NATIONALIST FORCES WITH THE COMMUNIST BLOC.

  IN RECENT YEARS THE CHINESE COMMUNISTS HAVE MAINTAINED THAT A NEW WORLD WAR WOULD RESULT IN THE UTTER DESTRUCTION OF THE IMPERIALIST CAMP AND THE COMPLETE COLLAPSE OF THE ENTIRE CAPITALIST SYSTEM.

  WORLD CONQUEST IN TRADITIONAL MILITARY TERMS AND WORLD REVOLUTION IN COMMUNIST TERMS ARE VERY DIFFERENT CONCEPTS.

  THE HISTORY OF THE ENTIRE COMMUNIST WORLD MOVEMENT HAS BEEN CHARACTERIZED BY TACTICAL OPPORTUNISM AND ADAPTABILITY, AND MAO TSE-TUNG, DURING THE LONG STRUGGLE FOR POWER IN CHINA, ELABORATED A DOCTRINE OF FLEXIBILITY WITH A UNIQUE CHINESE FLAVOR.

- THE CHINESE COMMUNISTS HAVE SHOWN NO INHIBITIONS ABOUT INVADING TIBET AND DOMINATING IT WITH OVERPOWERING FORCE, DESPITE STRONG OPPOSITION WITHIN TIBET AND HIGHLY UNFAVORABLE REACTIONS IN THE OUTSIDE WORLD, BECAUSE THEY REGARD TIBET AS CHINESE TERRITORY.

- SEVERAL OF CHINA'S FRONTIERS REMAIN TO BE FINALLY AND EXACTLY DEMARCATED, INCLUDING PORTIONS OF ITS FRONTIERS WITH INDIA, BURMA, AND OUTER MONGOLIA, AND SINCE 1949 THE CHINESE COMMUNISTS HAVE EXERTED OUTWARD PRESSURE AT SEVERAL POINTS ON CHINA'S PERIPHERY.

- THE CHINESE NATIONALIST REGIME, FOR EXAMPLE, DESPITE ITS RELATIVE INTERNAL WEAKNESS, ALSO CLAIMED TIBET, OUTER MONGOLIA, TAIWAN, AND CERTAIN TERRITORIES ON THE BORDERS OF BURMA AND INDIA, AND IT, TOO, ATTEMPTED TO EXERT ITS INFLUENCE AND DEVELOP SPECIAL RELATIONSHIPS, WITHIN THE LIMITS OF ITS CAPABILITIES, IN ALL THE AREAS ALONG CHINA'S PERIPHERY.

- TO STRENGTHEN THEIR SECURITY, EXPAND CHINESE INFLUENCE, AND PROMOTE THE SPREAD OF COMMUNISM, THE CHINESE COMMUNISTS HOPE, IN THE LONG RUN, TO FORCE THE WITHDRAWAL OF THE UNITED STATES AND OTHER WESTERN POWERS FROM ALL OF ASIA AND TO NEUTRALIZE THE NON-COMMUNIST NATIONS, PARTICULARLY JAPAN, IN THAT REGION.

- QUITE OBVIOUSLY, PEKING REGARDS THE UNITED STATES, CURRENTLY THE ONLY NON-COMMUNIST NATION STRONG ENOUGH TO COUNTERBALANCE COMMUNIST CHINA'S POWER IN ASIA, AS THE MAJOR OBSTACLE TO BOTH ITS SHORT-RANGE AND ITS LONG-RANGE AIMS AND, THEREFORE, AS ITS PARAMOUNT ENEMY.

  YET, THE CHINESE COMMUNISTS WILL CERTAINLY CONTINUE TO PRESS TOWARD THEIR LONG-RANGE REVOLUTIONARY AIMS, NOT ONLY FOR IDEOLOGICAL REASONS, BUT ALSO BECAUSE THE SPREAD OF COMMUNIST REGIMES ACROSS ASIA WOULD PROMOTE THE IMMEDIATE INTERESTS OF CHINA AS WELL AS THOSE OF THE WORLD REVOLUTIONARY MOVEMENT.

- CHINA'S CONCERNS REGARDING ITS DOMESTIC SITUATION, TERRITORIAL CLAIMS, AND NATIONAL SECURITY INTERESTS AS A MEMBER OF THE COMMUNIST BLOC, AND ITS DESIRE TO EXPAND CHINESE AND COMMUNIST INFLUENCE IN ASIA THROUGH TACTICS OF PEACEFUL COEXISTENCE DO NOT ALWAYS COINCIDE.

- IN THAT PERIOD PEKING, SEEMINGLY MOTIVATED IN LARGE PART BY ITS DESIRE TO SEAL OFF TIBET FROM THE OUTSIDE WORLD AS WELL AS TO ASSERT CERTAIN TRADITIONAL CHINESE TERRITORIAL CLAIMS, BROUGHT STRONG PRESSURES TO BEAR ON THE INDIAN BORDER.

- ITS POLICY, LIKE THAT OF OTHER NATIONS, IS SHAPED TO A CONSIDERABLE DEGREE BY A PROCESS OF ACTION AND INTERACTION BETWEEN ITSELF AND OTHER MAJOR POWERS, AND BETWEEN ITS LEADERS' AMBITIONS AND THE STUBBORN REALITIES OF THE OUTSIDE WORLD.

  FOR MANY YEARS TO COME, ASIA WILL BE CAUGHT UP IN A PROCESS OF PROFOUND CHANGE, AND COMMUNIST CHINA WILL EXPLOIT THIS TURBULENT PROCESS IN EVERY WAY IT CAN IN ORDER TO PROMOTE BOTH ITS OWN NATIONAL INTERESTS AND THOSE OF WORLD COMMUNISM.

- THE UNITED STATES IS UNAVOIDABLY ENGAGED IN A BASIC POWER STRUGGLE WITH COMMUNIST CHINA, AND THE CHINESE COMMUNISTS' ATTITUDE TOWARD THE USE OF MILITARY POWER MAKES IT ESSENTIAL TO BUILD UP THE NECESSARY MILITARY STRENGTH TO COUNTERBALANCE PEKING'S POWER.

Fig. 3a Extract Sentences: "Mao's Strategy" (Cont.)

COMMONLY, THE CELL WALL IS PICTURED AS SOMETHING OF A FINAL REPOSITORY FOR CERTAIN PRODUCTS OF CELLULAR METABOLISM.

DURING THE ACTIVE LIFE OF THE CELL, THE CHANGING PATTERN OF CELL WALL CONSTITUTION AND ORGANIZATION IS A REFLECTION OF THE CHANGING PHYSIOLOGICAL STATES AND BIOCHEMICAL CAPACITIES OF THE PROTOPLAST.

- OUR PRESENT CONSIDERATION OF CELL WALL DYNAMICS WILL CENTER ABOUT THREE AREAS: FIRST, CURRENT KNOWLEDGE AS TO THE BIOSYNTHESIS OF INDIVIDUAL CELL WALL COMPONENTS; SECOND, THE PATTERNS OF CHANGE IN CONSTITUTION AND ARCHITECTURE DURING GROWTH AND DIFFERENTIATION; AND THIRD, THE LITTLE KNOWLEDGE NOW AT HAND ABOUT CHEMICAL REGULATION OF CELL WALL FORMATION.

THE CHEMICAL UNIQUENESS OF LIGNINS AMONG WALL POLYMERS ALLOWS THEM TO BE DISTINGUISHED FROM THE MUCH INTERGRADED WALL POLYSACCHARIDES, AN ADVANTAGEOUS PROPERTY.

(C) TRANSMETHYLATION AS IN 3,4-DIHYDROXYPHENYLPYRUVIC ACID YIELDS (METHIONINE METHYLTRANSFERASE) 3-METHOXY-4-HYDROXYPHENYL (GUAIACYL) PYRUVIC ACID.

- OF PARTICULAR NOTE ARE SHIKIMIC ACID, A TRIHYDROXYLATED CYCLOHEXENE-CARBOXYLIC ACID AND PREPHENIC ACID, 1-CARBOXY-4-HYDROXY-CYCLOHEXADIENYL-PYRUVIC ACID.

- IN PROLIFERATING TOBACCO CALLUS TISSUE, SOME 25 PER CENT OF TOTAL CELL NITROGEN IS ASSOCIATED WITH THE INSOLUBLE CELL WALL FRACTION, WHEREAS NON-PROLIFERATING PITH WALLS CONTAIN BUT 2 PER CENT OF THE CELLULAR NITROGEN.

- THE VARIOUS BIOCHEMICAL THEORIES OF LIGNIFICATION FALL CONVENIENTLY INTO THREE CATEGORIES: (A) LIGNIN ARISES IN THE CELL WALL BY DIRECT TRANSFORMATION OF OTHER WALL COMPONENTS, (B) LIGNIN ARISES FROM PRECURSORS WHICH DIFFUSE CENTRIPETALLY FROM THEIR POINT OF ORIGIN IN THE CAMBIUM AND BECOME INCORPORATED INTO THE WALLS OF XYLEM AND PHLOEM, AND (C) LIGNIN ARISES FROM CYTOPLASMIC PRECURSORS FORMED IN DIFFERENTIATING CELL AND IS SUBSEQUENTLY INCORPORATED IN THE WALL.

- IT HAS ALSO BEEN SUGGESTED THAT THE PIT FIELDS OF THE WALL SERVE AS CENTERS OF CELLULOSE SYNTHESIS, BUT THE RECENT CONCEPT OF MULTINET GROWTH SEEMS TO BE OF CONSIDERABLE VALUE IN RELATING WALL SYNTHESIS WITH CELL EXTENSION.

THE SYNTHESIS OF VARIOUS CELL WALL COMPONENTS UNDER THE INFLUENCE OF 3-INDOLEACETIC ACID OR OTHER AGENTS HAS BEEN STUDIED IN SEVERAL TISSUES.

- FUNCTIONALLY, CELL WALL LYSIS MAY BE ASSOCIATED WITH SEVERAL KINDS OF BIOLOGICAL BEHAVIOR OR INTERACTIONS AMONG ORGANISMS. (A) RESORPTION OF FORMED WALL STRUCTURES OCCURS DURING GROWTH AND DEVELOPMENT.

LYSOZYME HEXOSAMINE-PEPTIDE CELL WALL YIELDS SOLUBLE HEXOSAMINE-PEPTIDES AND HEXOSAMINE-MURAMIC ACID DIMER (BACTERIA, MAMMALS).

IN CONTRAST TO A DIRECT TEST OF THE PROPOSITION WHICH IMPLICATES A SPECIFIC ENZYME, THE POSSIBLE PHYSICAL PARTICIPATION OF THE CELL WALL CAN BE TESTED MORE DIRECTLY, BY RECOURSE TO A SUITABLE MODEL FOR CELL WALL SUBSTANCE.

CELLULOSE EXHIBITS ONE-TWENTIETH AND CHITIN ONE-SEVENTH THE ACTIVITY MEASURED IN CELL WALL SYSTEMS, WHEREAS PECTIC ACID AND METHYL CELLULOSE WERE FOUND TO EXCEED CELL WALL PREPARATIONS 2 = 10-FOLD.

HENCE, THE NET ACTIVITY OF THE CELL WALL MAY INVOLVE A COMPARATIVELY SMALL CONTRIBUTION FROM CELLULOSE ITSELF TOGETHER WITH LARGE CONTRIBUTIONS FROM POLYSACCHARIDES OF LOWER MOLECULAR WEIGHT SUCH AS PECTIC ACID.

Fig. 3b Extract Sentences: "Cell Wall Dynamics"

selection percents. Analysis of those tests showed selection percents for the four points varying widely from 0.2% to 12.5%, coselection percents from 0 to 100%, and acceptable selection percents from 75 to 100%. It revealed that the algorithm had done a good job and had picked the best or next best cutoff point. Points 1, 2, and 3 showed great variability in usefulness, but for these four texts, point 4 did not. It was chosen by the algorithm in two of the four texts and was as good as the algorithmic choice for the other two. Because point 4 is the most economical to calculate, this is an interesting result. Since tests were made on only four texts, point 4 cannot be regarded as dependable, but certainly it would merit further investigation.

2.3 THE FREQUENCY-SYNTAX INDEXING EXPERIMENT

When an algorithm for indexing using both frequency and syntactic criteria was developed, it was found that many of the techniques used in extracting (see Ref. 3 and section 1.2 of this report) could be applied to indexing, but with modifications. Again it was found useful to reduce the text to all nonfunction words of all noun phrases in the text. This time, however, it was necessary to keep the noun phrases intact, since it was believed that it would be such noun phrases which would be useful for an index. For purposes of the frequency count, the noun phrases were broken up into their constituent words as before. Once the frequencies are calculated, one is again faced with the problem of the frequency threshold below which the words (and the phrases from which they came) are to be discarded. The criteria used in the extract experiments cannot be used here, because many more words should be included in the index than were useful in choosing sentences. It was decided to try two threshold points, the first to include down to the frequency corresponding to 40% of the total number of words in the reduced text and the second to include down to the frequency corresponding to 10% of the number of unique words in the reduced text. (The first threshold is like that used in point 1 of the extract experiment with V1 set at 40.) The following steps were thus necessary for the formation of the index:

(1) All the nonfunction words in noun phrases were extracted from the text and stored with the noun phrase from which they came, and with the text page and sentence number from which they came. A partial list of such words and phrases, after alphabetization, is shown in Table 4.

Table 4

EXCERPT FROM THE ALPHABETIZED LIST OF WORDS AND PHRASES EXTRACTED FROM ART, LIFE AND EXPERIMENT

(The 6 digit hexadecimal number gives the text page and sentence number)

SEVENTEENTH	00D001 SEVENTEENTH CENTURY
SEVENTEENTH-CENTURY	00D001 SEVENTEENTH-CENTURY ENGLAND
SHADOWS	00E008 SHADOWS
SHALLOW	00CF02 SHALLOW GALLERIES
SHAPE	00E105 SHAPE
SHEETS	00CF04 SHEETS
SHIPWRECKED	00E306 SHIPWRECKED EUROPEANS
SHOCK	00D307 SHOCK
SHOWED	00D306 SHOWED VESSELS
SHOWN	00E808 HAVING SHOWN
SHOWN	00E800 SHOWN
SHREWDLY	00D30A SHREWDLY
SHRIEK	00E801 SHRIEK
SIDE	00E900 ONE SIDE
SIDE	00D509 ONE SIDE
SIDE	00E00B POSITIVE SIDE
SIDE	00CD03 SIDE
SIDE	00C003 SIDE
SIDE	00D20A SIDE
SIDE	00E600 SIDE
SIDES	00CF02 SIDES
SIGHT	00E30B SIGHT
SIGNIFICANCE	00D50E SIGNIFICANCE
SIMPLE	00E202 SIMPLE
SIMPLE	00DB0F SIMPLE
SIMPLE	00D20E SIMPLE POINT
SIMPLE	00E50D SIMPLE TECHNIQUES
SIMPLE	00D405 SIMPLE WORD
SIMPLY	00D09 SIMPLY
SINGLE	00CF06 SINGLE DETAIL
SINGLE	00DB03 SINGLE PHENOMENON
SITUATES	00E305 SITUATES
SIX	00CE06 SIX YEARS
SIX	00CE0A SIX YEARS
SIXTEENTH	00CC09 SIXTEENTH
SKELETON	00D00B SKELETON
SKILL	00CF06 COMPREHENSIVE SKILL
SKIN	00D00B SKIN
SLAB	00CF03 SLAB
SLIPPERY	00D80D SLIPPERY
SLOW	00E800 SLOW HEART-BEAT
SMALLER	00DC0D SMALLER FORCES
SNOBBISH	00DD01 SNOBBISH
SNOW	00D10D SNOW
SOCIETIES	00DD00 SOCIETIES
SOCIETY	00E308 SOCIETY
SOLE	00D909 SOLE
SOMEWHAT	00CD0D SOMEWHAT
SORCERY	00E201 SORCERY
SORT	00D909 CIRCULAR SORT
SOURCE	00CC0B SOURCE
SOURCE	00CD01 SOURCE
SOURCE	00D003 SOURCE
SOURCE	00D80A SOURCE
SOUTHWEST	00CB05 SOUTHWEST
SOVEREIGN	00E307 SOVEREIGN
SOVEREIGNTY	00D204 SOVEREIGNTY
SPAN	00D603 CHRONOLOGICAL SPAN
SPANIARD	00D806 SPANIARD
SPECIAL	00D000 SPECIAL APPEAL
SPECIAL	00DA0B SPECIAL ORGANS
SPECIALTIES	00DE03 SPECIALTIES
SPECIES	00CD02 SPECIES
SPECIOUS	00D404 SPECIOUS VIEWS
SPECULATIONS	00E006 SPECULATIONS
SPIRIT	00D501 HOLY SPIRIT
SPIRIT	00CC0C SPIRIT
SPIRIT	00CD0F SPIRIT
SPIRIT	00CF0D SPIRIT
SPIRIT	00D20D SPIRIT
SPIRIT	00E809 VITAL SPIRIT
SPIRIT	00E20D ANIMAL SPIRITS
SPIRITS	00D503 THREE SPIRITS
SPIRITS	00D20B VITAL SPIRITS
SPIRITS	00E20D VITAL SPIRITS
SPIRITS	00D509 VITAL SPIRITS
SQUEAMISH	00CE0F SQUEAMISH
STAGE	00DE05 STAGE PLAYS
STAGES	00E00B STAGES
STAKE	00D40A STAKE
STAND	00CF02 STAND
STANDARD	00E304 AMERICAN STANDARD
STANDING	00D00D STANDING
STANDING	00E007 STANDING
STANDING	00CF0D STANDING RULE
START	00E00C START ANEW
STATE	00D601 STATE
STATE	00DA0D STATE
STATE	00E300 STATE
STATEMENT	00D105 STATEMENT
STATURE	00CD0E COMPARABLE STATURE
STATURE	00DC0C STATURE
STEP	00CD05 FIRST STEP
STEP	00DF01 FIRST STEP
STEPS	00E00E SUCCESSIVE STEPS
STERILE	00DF03 STERILE HABIT
STOMACH	00E207 STOMACH
STREAM	00DA01 BLOOD STREAM
STREAM	00DB13 BLOOD STREAM
STRENGTH	00DC0D STRENGTH
STREWN	00CC02 SUPERSTITION STREWN
STROKE	00D209 EXPANDING STROKE
STROKE	00D800 PUMPING STROKE
STRUCTURE	00D304 ENTIRE STRUCTURE
STRUCTURE	00CC03 EVOLUTIONARY STRUCTURE
STRUCTURE	00CB03 PHYSICAL STRUCTURE
STRUCTURE	00D00A STRUCTURE
STRUCTURE	00D00A STRUCTURE
STRUCTURE	00D504 STRUCTURE
STRUCTURES	00DA00 CAPILARY STRUCTURES
STRUGGLES	00E10E STRUGGLES
STUDENT	00DE04 ACUTE STUDENT
STUDENT	00D60B MEDICAL STUDENT
STUDENT	00CF0D STUDENT
STUDENT	00DE04 STUDENT
STUDENTS	00CF0D MEDICAL STUDENTS
STUDENTS	00D700 PADUAN STUDENTS
STUDENTS	00CF03 STUDENTS
STUDIES	00DB0E ORGANIC STUDIES
STUDIES	00CE03 STUDIES
STUDY	00D00A STUDY
STUFFING	00EF0D STUFFING
STUNNING	00CB02 STUNNING WOODCUTS
STUPID	00D806 STUPID ERROR
STYLE	00CB07 STYLE
STYLE	00E108 STYLE
SUBJECT	00CF0B SUBJECT
SUBJECT	00CF03 SUBJECT
SUBJECT	00D107 SUBJECT
SUBJECT	00D303 SUBJECT
SUBJECT	00DC01 SUBJECT
SUBJECT	00D706 SUBJECT
SUBJECT	00DB04 SUBJECT
SUBJECT	00DB0A SUBJECT
SUBJECT	00D006 SUBJECT
SUBJECT	00CB00 SUBJECT MATTER
SUBJECT	00DE00 SUBJECT MATTER
SUBJECT	00CB0B VESALIUS'S SUBJECT
SUBJECT	00DA01 PARTICULAR SUBJECTS
SUBJECT	00CC09 SUBJECTS
SUBJECT	00CC09 SUBJECTS
SUBORDINATE	00CB0A SUBORDINATE PLACE
SUBSIDY	00E20A SUBSIDY
SUBSTANCE	00CE0C SUBSTANCE
SUBSTANCE	00D000 SUBSTANCE
SUBTLE	00D505 SUBTLE BLOOD
SUBTLETIES	00E206 SCHOLASTIC SUBTLETIES
SUCCESS	00CE03 GREAT SUCCESS
SUCCESS	00CF04 SUCCESS
SUCCESS	00D701 SUCCESS
SUCCESSION	00E20D SUCCESSION
SUCCESSIVE	00E00E SUCCESSIVE STEPS
SUCCESSOR	00D50B VESULIUS'S SUCCESSOR
SUCCESSORS	00CE02 ALARMED SUCCESSORS
SUCTION	00D105 SUCTION
SUGGESTIVE	00D50A SUGGESTIVE
SUIT	00CD08 SUIT
SUM	00D909 SUM
SUN	00D202 SUN
SUN	00DC03 SUN
SUN-WORSHIPER	00DA07 SUN-WORSHIPER
SUPER-INDUCING	00E001 SUPER-INDUCING
SUPERIORITY	00CD06 SUPERIORITY
SUPERSTITION	00CC01 SUPERSTITION STREWN
SUPPORTER	00D30B SUPPORTER YEARS
SUPPOSE	00DC0C SUPPOSE
SUPPOSE	00E101 SUPPOSE
SUPPOSITIONS	00EF06 GRATUITOUS SUPPOSITIONS
SUPPOSITIONS	00D905 SUPPOSITIONS
SURGEON	00D009 SURGEON
SURGERY	00CF01 SURGERY
SURGERY	00D703 SURGERY
SURNAME	00CE00 SURNAME
SUSPECT	00DF0B SUSPECT
SYMBOL	00D003 SYMBOL
SYNCHRONIZES	00D801 SYNCHRONIZES
SYNONYM	00E109 SYNONYM
SYSTEM	00D30B GALENIC SYSTEM
SYSTEM	00D00D MUSCULAR SYSTEM
SYSTEM	00D00B NERVOUS SYSTEM
SYSTEM	00D009 SYSTEM
SYSTEM	00DF06 SYSTEM
SYSTEM	00DF08 SYSTEM
SYSTEM	00DF0C SYSTEM
SYSTEM	00D00B VASCULAR SYSTEM
SYSTEMATIC	00CF05 SYSTEMATIC APPROACH
SYSTEMATIC	00DF05 SYSTEMATIC DOGMAS
SYSTEMATIC	00DB0F SYSTEMATIC EXTENSION
SYSTEMATIC	00D607 SYSTEMATIC THOUGHT
SYSTEMATIC	00D909 SYSTEMATIC WORK
SYSTEMS	00DF05 RECEIVED SYSTEMS
SYSTOLE	00D305 SYSTOLE
TABLE	00E003 TABLE
TABULATIONS	00CB02 FORBIDDING TRIGONOMETRICAL TABULATIONS
TACTICS	00CF0A TACTICS
TAKES	00E10E VEXED TAKES
TARGET	00E110 LARGEST TARGET
TAXONOMY	00CD09 TAXONOMY
TEACHER	00CF03 TEACHER
TEACHERS	00CF01 TEACHERS
TEACHING	00CE07 TEACHING
TEACHING	00E004 TEACHING
TEACHING	00D80B TEACHING
TECHNICAL	00E407 TECHNICAL TRADITION

(2) The list of words was alphabetized, and a frequency count of the words was made.

(3) The frequency counts were then ordered (descending order).

(4) The percent of each frequency count of the total of all frequency counts was calculated and stored with the cumulative percent, up to the point at which the cumulative frequency was greater than 40. This frequency was the first threshold point, called VALUEA.

(5) The frequency corresponding to 10% of the number of total entries in the frequency table was calculated. This frequency was the second threshold point, called VALUEB.

(6) The words and phrases corresponding to all frequencies down to VALUEA were located and printed with their corresponding page and sentence numbers.

(7) Step (6) was repeated for VALUEB instead of VALUEA.

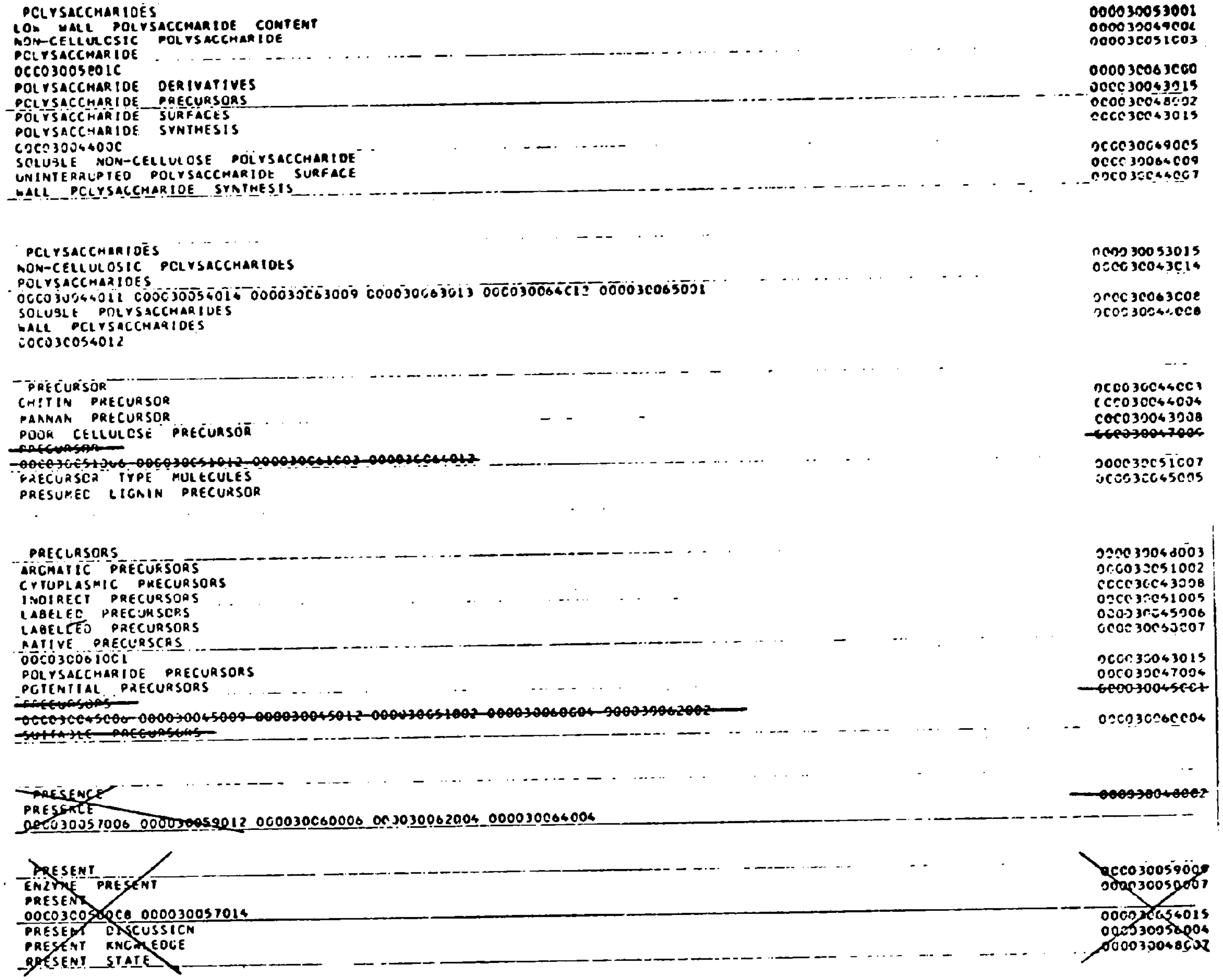
The words were printed in alphabetical order and under each word were the phrases containing that word, also in alphabetical order. Each phrase was followed bv the sentence and page number or numbers of all instances of that phrase in the text. If a phrase has more than one page reference, these references will appear on the next line.

To test the algorithm, a chapter from each of six texts (1, 3, 4, 7, 8, and 9 of Ref. 7) was indexed by the program INDEX. Before evaluating the indexes, they were hand-edited to remove phrases or sometimes whole word and phrase groups which were obviously low in information content. In all six texts, the index produced using the threshold VALUEB was used because it was most complete, although it also required the most editing. For the chapter "Roots of Mao's Strategy" (from 1 of Ref. 7) there were 95 high-frequency words included using the VALUEB threshold. Of these, 35 were edited out; a quick glance is usually all that is necessary to show that the phrases subsumed under these words are low in information content. Words edited out included such as "future, long, major, position, threat, years." Of the 60 remaining items, 34 were missing from the index produced using VALUEA as the threshold. Words missing included such as "bloc, economic, ideological, Mao, Peking, socialist, Tibet."

An excerpt from the index printout for "Cell Wall Dynamics" is given in Fig. 4. Phrases which have been edited out have a single line drawn through them. Words which have been eliminated with all their phrases are crossed out. The phrase listings are interrupted by numbers because, as explained above, if a phrase has more than cne page reference, the pages are listed on the next line.

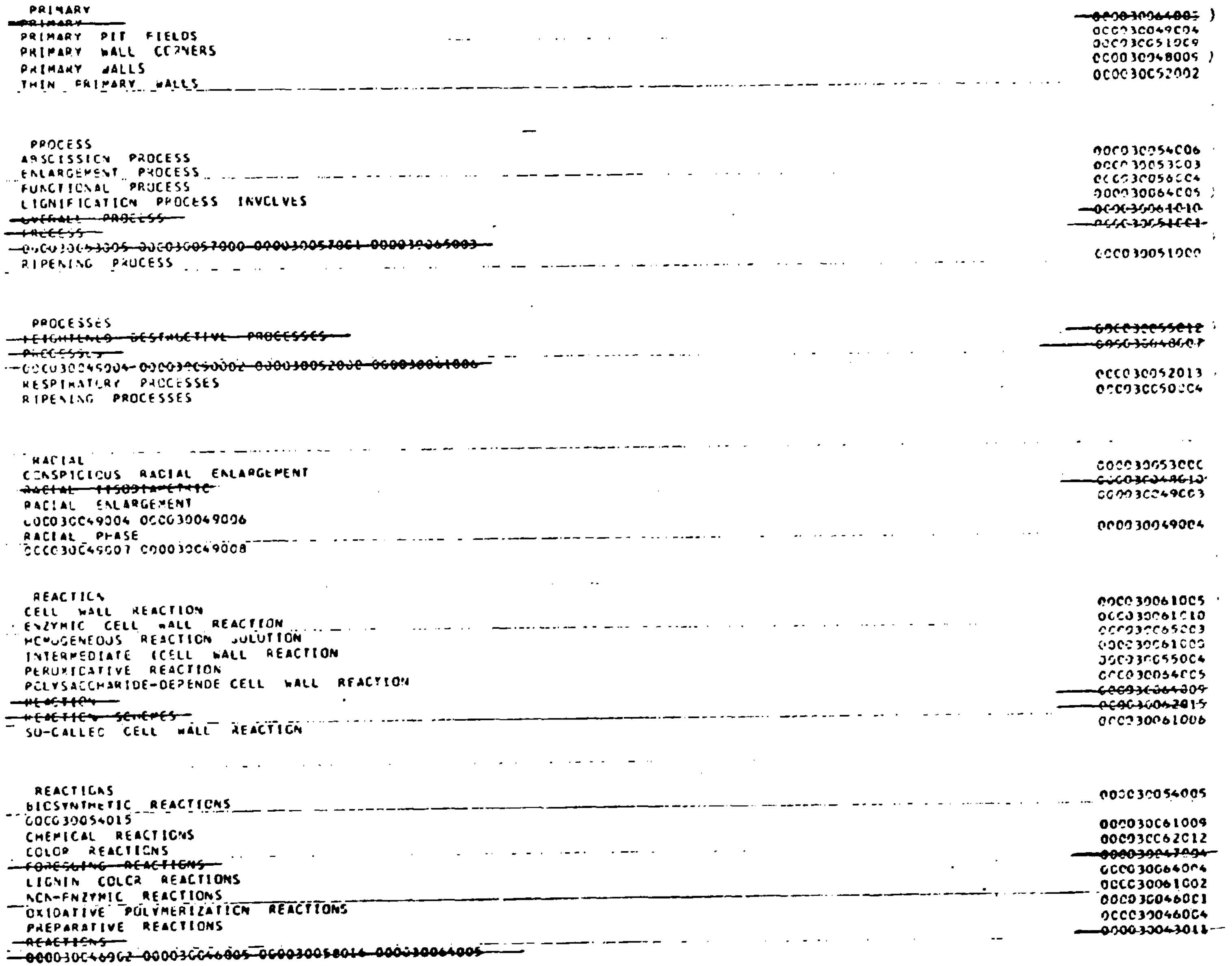
Evaluating the results of automatic indexing is always difficult. It is not the sort of thing which can be objectively and precisely measured. Nevertheless, it is possible to get a good feel for the usefulness and adequacy of coverage of an index. Since they were better, only the indexes produced with VALUEB as threshold were evaluated. After the automatic index of a chapter was edited, the remaining words and their phrases were compared with the index items for that chapter in the back of the book. Generally, it was the phrases of the automatic index and not the single words which provided useful information. For example, "concepts" by itself is not very useful, but "basic Marxist concepts" is. Names, such as "Japan," are the exception to this.

For each text, the items were counted which were found both in the automatic and the back-of-the-book index. Remaining items in the back of the book were then counted, and also those remaining items in the automatic index which seemed especially important to the author. These statistics are helpful in evaluating the automatic index and are shown in Fig. 5. Along with them the number of total words and phrases in the automatic index are given so that the results can be kept in perspective. The figures given for the "Ceil Wall Dynamics" are for the regular index only. In the back of the book there is also a special index of organisms. There were 26 items in this index, only 1 of which appeared in the automatic index. The organisms were usually mentioned just once and were not integral to the subject matter, so it is difficult to see how an algorithm could be designed to pick them out.



```
 PCLYSACCHARIDES                                           000030053001
LOW  WALL  POLYSACCHARIDE  CONTENT                         000030049006
NON-CELLULCSIC  POLYSACCHARIDE                             000030051003
PCLYSACCHARIDE
0CC03005801C                                               000030063C00
POLYSACCHARIDE  DERIVATIVES                                000030043015
PCLYSACCHARIDE  PRECURSORS                                 0C0030048002
POLYSACCHARIDE  SURFACES                                   0CCC30043015
POLYSACCHARIDE  SYNTHESIS
C0C030044000                                               0CC030049005
SOLUBLE  NON-CELLULOSE  POLYSACCHARIDE                     000C30064009
UNINTERRUPTED  POLYSACCHARIDE  SURFACE                     000030044007
WALL  PCLYSACCHARIDE  SYNTHESIS

 PCLYSACCHARIDES                                           000030053015
NON-CELLULOSIC  PCLYSACCHARIDES                            0CCC30043C14
POLYSACCHARIDES
0CC030044011 C00C30054014 000030C63009 C00030063013 00C030064C12 000030065001   0C0C30063C08
SOLUBLE  POLYSACCHARIDES                                   0C0C30044008
WALL  PCLYSACCHARIDES
C0C03C054012

 PRECURSOR                                                 0C0030044CC1
CHITIN  PRECURSOR                                          CCC030044004
MANNAN  PRECURSOR                                          C0C030043008
POOR  CELLULCSE  PRECURSOR                                 ~~000030047006~~
~~PRECURSOR~~
~~00C030051006 000030051012 000030061003 000030064012~~    000030051C07
PRECURSOR  TYPE  MOLECULES                                 0C0030045005
PRESUMEC  LIGNIN  PRECURSOR

 PRECURSORS                                                000030048003
AROMATIC  PRECURSORS                                       0C0030051002
CYTOPLASMIC  PRECURSORS                                    CCCC30043008
INDIRECT  PRECURSORS                                       000030051005
LABELEC  PRECURSORS                                        000030045006
LABELLED  PRECURSORS                                       0000300500C7
NATIVE  PRECURSCRS
00C030061001                                               0CCC30043015
POLYSACCHARIDE  PRECURSORS                                 00C030047004
PCTENTIAL  PRECURSORS                                      ~~0C0030045C04~~
~~PRECURSORS~~
~~0CC030045006 000030045009 000030045012 000030051002 000030060004 000030062002~~   000030060004
~~SUITABLE  PRECURSORS~~

~~PRESENCE~~                                               ~~000030048002~~
PRESENCE
00C030057006 000030059012 0C0030C60006 000030062004 000030064004

 PRESENT                                                   0CC030059009
ENZYME  PRESENT                                            000030050007
PRESENT
00C030050008 000030057014                                  000030054015
PRESENT  DISCUSSICN                                        000030056004
PRESENT  KNOWLEDGE                                         000030048007
PRESENT  STATE
```

Fig. 4 Sample of Index Printout From "Cell Wall Dynamics" Showing Editing



PRIMARY	
~~PRIMARY~~	~~000030064005~~)
PRIMARY PIT FIELDS	000030049004
PRIMARY WALL CORNERS	000030051009
PRIMARY WALLS	000030048009)
THIN PRIMARY WALLS	000030052002
PROCESS	
ABSCISSION PROCESS	000030054006
ENLARGEMENT PROCESS	000030053003
FUNCTIONAL PROCESS	000030056004
LIGNIFICATION PROCESS INVOLVES	000030064005
~~OVERALL PROCESS~~	~~000030061010~~
~~PROCESS~~	~~000030051001~~
~~000030053005 000030057000 000030057001 000030065003~~	
RIPENING PROCESS	000030051000
PROCESSES	
~~HEIGHTENED DESTRUCTIVE PROCESSES~~	~~000030055012~~
~~PROCESSES~~	~~000030048007~~
~~000030045004 000030050002 000030052000 000030061006~~	
RESPIRATORY PROCESSES	000030052013
RIPENING PROCESSES	000030050004
RACIAL	
CONSPICUOUS RACIAL ENLARGEMENT	000030053000
~~RACIAL ISODIAMETRIC~~	~~000030049010~~
RACIAL ENLARGEMENT	000030049003
000030049004 000030049006	
RACIAL PHASE	000030049004
000030049007 000030049008	
REACTION	
CELL WALL REACTION	000030061005
ENZYMIC CELL WALL REACTION	000030061010
HOMOGENEOUS REACTION SOLUTION	000030065003
INTERMEDIATE (CELL WALL REACTION	000030061003
PEROXIDATIVE REACTION	000030055004
POLYSACCHARIDE-DEPENDE CELL WALL REACTION	000030054005
~~REACTION~~	~~000030064009~~
~~REACTION SCHEMES~~	~~000030062015~~
SO-CALLED CELL WALL REACTION	000030061006
REACTIONS	
BIOSYNTHETIC REACTIONS	000030054005
000030054015	
CHEMICAL REACTIONS	000030061009
COLOR REACTIONS	000030062012
~~FOREGOING REACTIONS~~	~~000030047004~~
LIGNIN COLOR REACTIONS	000030064004
NON-ENZYMIC REACTIONS	000030061002
OXIDATIVE POLYMERIZATION REACTIONS	000030046001
PREPARATIVE REACTIONS	000030046004
~~REACTIONS~~	~~000030043011~~
~~000030046902 000030046005 000030058016 000030064005~~	

Fig. 4 Sample of Index Printout From "Cell Wall Dynamics" Showing Editing (Cont.)

Text Chapter Name	RMS	ETR	OC	BPG	ALE	CWD
Number items back-of-the-book only	30	3	31	17	59	43
Number items automatic only	36	13	17	30	48	163
Number items common to both	19	8	34	33	38	50
Total number words selected	59	24	49	28	51	76
Total number items automatic index	355	80	155	235	133	411

RMS	=	Roots of Mao's Strategy	BPG	=	Basic Public Goals
ETR	=	Einstein's Theory of Relativity	ALE	=	Art, Life, and Experiment
OC	=	Occupation and Careers	CWD	=	Cell Wall Dynamics

Fig. 5 Comparison of Back-of-the-Book and Automatic Indexes

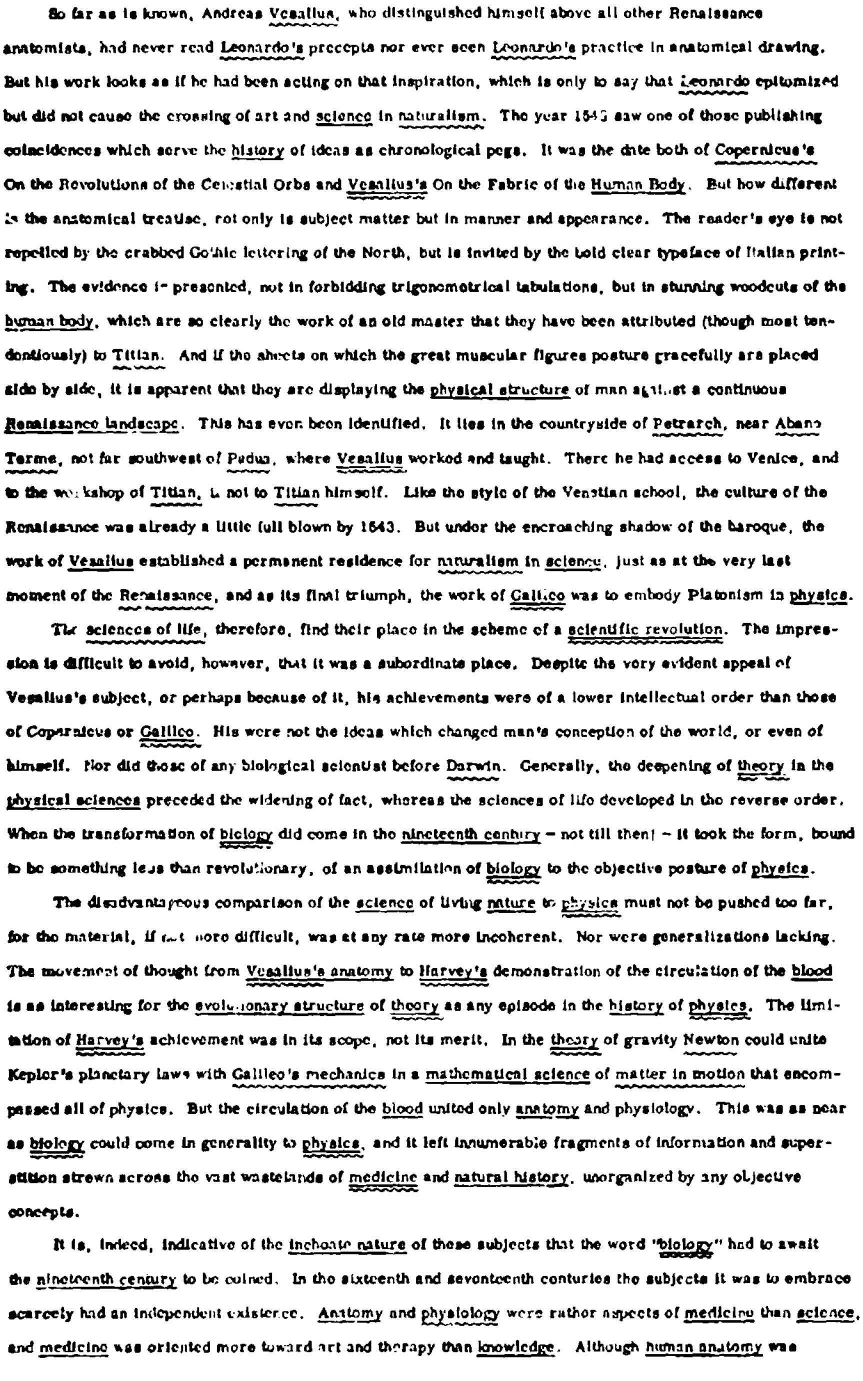
It is clear from Fig. 5 that the automatic index produces many more index items, roughly from 1.5 to 7 times as many as in the back-of-the-book indexes. This is partly because each item is likely to occur more than once under different words. Thus "cellulose synthesis" will occur under "cellulose" and also under "synthesis." This occurs to a lesser extent in back-of-the-book indexes also and is more of a convenience than a drawback. Another factor is that an important word will pick up any number of interesting though not always necessary items. For example, in "Roots of Mao's Strategy," the word <u>Chinese</u> picked up 22 different noun phrases, among them "Chinese national interests, Chinese communist approach, Chinese territorial conquest, top Chinese communist leaders," etc.

Study of Fig. 5 will show that the automatic index picked up from 40 to 75% of the items in the back-of-the-book index (four of them 40 to 55% and two from 65 to 75%). In addition, it picked up many good items which were not in the back-of-the-book indexes. It has an advantage over the back-of-the-book index in that it never misses a reference page. For example, in "The Roots of Mao's Strategy," the back of the book indicates that "United States" is referenced on pages 82—83; the automatic index shows it references on pages 73, 78—83 and gives the sentence numbers on each page.

Perhaps the best way to show the kind of items provided by the automatic index and how they compare with those in the back-of-the-book index is to show text with the items referenced by the indexes. This has been done for a portion of "Art, Life, and Experiment," Fig. 6. Index items from the back of the book are underscored by a wavy line, and those from the automatic index are underscored by a solid line. Note that only those proper names which are mentioned with high frequency have been included in the program-chosen index terms. Thus Vesalius, Galileo, and Harvey are indexed by bot but Leonardo, Copernicus, Titian, Darwin, Newton, Petrarch, Abano, Terme, Padua, and Charles V are indexed only in the back of book. In a book of this kind, either all capitalized words should be included, or all names in a famous name dictionary should be included.

It is the author's feeling that these results are very encouraging. There is every indication that satisfactory back-of-the-book indexes could be produced automatically, with a human editor to cut out superfluous items. In fact, considering the variability in quality of human-compiled indexes, the present algorithm produces a useable if rather voluminous index. An addition of an algorithm to select proper names regardless of frequency would be an improvement. So also would the equating of words with similar stems in the calculation of the frequency of occurrence of words, and this will be the subject of the next experiment. It will also be interesting, though of more uncertain result, to vary the syntactic criteria by which the text is reduced before frequency calculations are made.

One very important potential use of a computer-produced index such as this is in automatic retrieval. The Information Sciences Laboratory (Ref. 8) has developed an information retrieval system called DIALOG in which search is conducted by logical combination of index items, with display of alphabetically near items to facilitate item selection. The form of the index produced by this algorithm (Fig. 4) would seem to be especially adaptable to this system. Thus the use of these indexes in automatic retrieval will be a subject for future research.



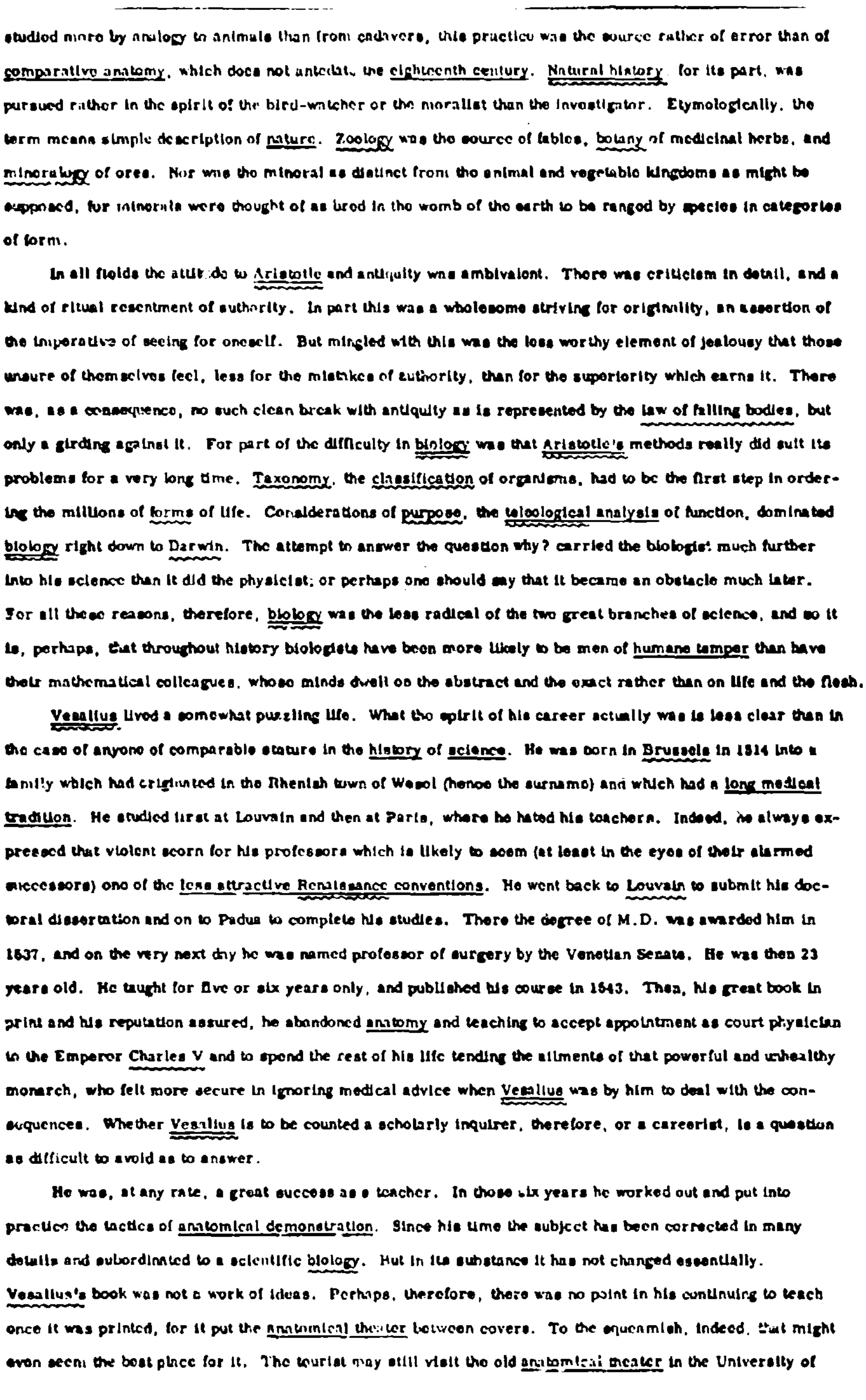
So far as is known, Andreas Vesalius, who distinguished himself above all other Renaissance anatomists, had never read Leonardo's precepts nor ever seen Leonardo's practice in anatomical drawing. But his work looks as if he had been acting on that inspiration, which is only to say that Leonardo epitomized but did not cause the crossing of art and science in naturalism. The year 1543 saw one of those publishing coincidences which serve the history of ideas as chronological pegs. It was the date both of Copernicus's On the Revolutions of the Celestial Orbs and Vesalius's On the Fabric of the Human Body. But how different is the anatomical treatise, not only in subject matter but in manner and appearance. The reader's eye is not repelled by the crabbed Gothic lettering of the North, but is invited by the bold clear typeface of Italian printing. The evidence is presented, not in forbidding trigonometrical tabulations, but in stunning woodcuts of the human body, which are so clearly the work of an old master that they have been attributed (though most tendentiously) to Titian. And if the sheets on which the great muscular figures posture gracefully are placed side by side, it is apparent that they are displaying the physical structure of man against a continuous Renaissance landscape. This has even been identified. It lies in the countryside of Petrarch, near Abano Terme, not far southwest of Padua, where Vesalius worked and taught. There he had access to Venice, and to the workshop of Titian, if not to Titian himself. Like the style of the Venetian school, the culture of the Renaissance was already a little full blown by 1543. But under the encroaching shadow of the baroque, the work of Vesalius established a permanent residence for naturalism in science, just as at the very last moment of the Renaissance, and as its final triumph, the work of Galileo was to embody Platonism in physics.

The sciences of life, therefore, find their place in the scheme of a scientific revolution. The impression is difficult to avoid, however, that it was a subordinate place. Despite the very evident appeal of Vesalius's subject, or perhaps because of it, his achievements were of a lower intellectual order than those of Copernicus or Galileo. His were not the ideas which changed man's conception of the world, or even of himself. Nor did those of any biological scientist before Darwin. Generally, the deepening of theory in the physical sciences preceded the widening of fact, whereas the sciences of life developed in the reverse order. When the transformation of biology did come in the nineteenth century – not till then! – it took the form, bound to be something less than revolutionary, of an assimilation of biology to the objective posture of physics.

The disadvantageous comparison of the science of living nature to physics must not be pushed too far, for the material, if not more difficult, was at any rate more incoherent. Nor were generalizations lacking. The movement of thought from Vesalius's anatomy to Harvey's demonstration of the circulation of the blood is as interesting for the evolutionary structure of theory as any episode in the history of physics. The limitation of Harvey's achievement was in its scope, not its merit. In the theory of gravity Newton could unite Kepler's planetary laws with Galileo's mechanics in a mathematical science of matter in motion that encompassed all of physics. But the circulation of the blood united only anatomy and physiology. This was as near as biology could come in generality to physics, and it left innumerable fragments of information and superstition strewn across the vast wastelands of medicine and natural history, unorganized by any objective concepts.

It is, indeed, indicative of the inchoate nature of these subjects that the word "biology" had to await the nineteenth century to be coined. In the sixteenth and seventeenth centuries the subjects it was to embrace scarcely had an independent existence. Anatomy and physiology were rather aspects of medicine than science, and medicine was oriented more toward art and therapy than knowledge. Although human anatomy was

Fig. 6 "Art, Life, and Experiment" Text Excerp+ Showing Back-of-Book and Program-Chosen Index Terms



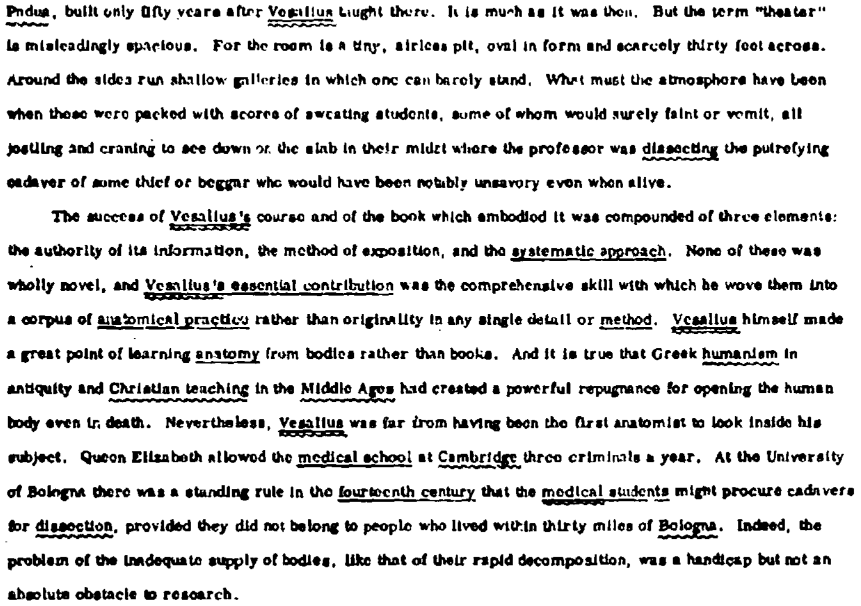
studied more by analogy to animals than from cadavers, this practice was the source rather of error than of comparative anatomy, which does not antedate the eighteenth century. Natural history, for its part, was pursued rather in the spirit of the bird-watcher or the moralist than the investigator. Etymologically, the term means simple description of nature. Zoology was the source of fables, botany of medicinal herbs, and mineralogy of ores. Nor was the mineral as distinct from the animal and vegetable kingdoms as might be supposed, for minerals were thought of as bred in the womb of the earth to be ranged by species in categories of form.

In all fields the attitude to Aristotle and antiquity was ambivalent. There was criticism in detail, and a kind of ritual resentment of authority. In part this was a wholesome striving for originality, an assertion of the imperative of seeing for oneself. But mingled with this was the less worthy element of jealousy that those unsure of themselves feel, less for the mistakes of authority, than for the superiority which earns it. There was, as a consequence, no such clean break with antiquity as is represented by the law of falling bodies, but only a girding against it. For part of the difficulty in biology was that Aristotle's methods really did suit its problems for a very long time. Taxonomy, the classification of organisms, had to be the first step in ordering the millions of forms of life. Considerations of purpose, the teleological analysis of function, dominated biology right down to Darwin. The attempt to answer the question why? carried the biologist much further into his science than it did the physicist; or perhaps one should say that it became an obstacle much later. For all these reasons, therefore, biology was the less radical of the two great branches of science, and so it is, perhaps, that throughout history biologists have been more likely to be men of humane temper than have their mathematical colleagues, whose minds dwell on the abstract and the exact rather than on life and the flesh.

Vesalius lived a somewhat puzzling life. What the spirit of his career actually was is less clear than in the case of anyone of comparable stature in the history of science. He was born in Brussels in 1514 into a family which had originated in the Rhenish town of Wesel (hence the surname) and which had a long medical tradition. He studied first at Louvain and then at Paris, where he hated his teachers. Indeed, he always expressed that violent scorn for his professors which is likely to seem (at least in the eyes of their alarmed successors) one of the less attractive Renaissance conventions. He went back to Louvain to submit his doctoral dissertation and on to Padua to complete his studies. There the degree of M.D. was awarded him in 1537, and on the very next day he was named professor of surgery by the Venetian Senate. He was then 23 years old. He taught for five or six years only, and published his course in 1543. Then, his great book in print and his reputation assured, he abandoned anatomy and teaching to accept appointment as court physician to the Emperor Charles V and to spend the rest of his life tending the ailments of that powerful and unhealthy monarch, who felt more secure in ignoring medical advice when Vesalius was by him to deal with the consequences. Whether Vesalius is to be counted a scholarly inquirer, therefore, or a careerist, is a question as difficult to avoid as to answer.

He was, at any rate, a great success as a teacher. In those six years he worked out and put into practice the tactics of anatomical demonstration. Since his time the subject has been corrected in many details and subordinated to a scientific biology. But in its substance it has not changed essentially. Vesalius's book was not a work of ideas. Perhaps, therefore, there was no point in his continuing to teach once it was printed, for it put the anatomical theater between covers. To the squeamish, indeed, that might even seem the best place for it. The tourist may still visit the old anatomical theater in the University of

Fig. 6 "Art, Life, and Experiment" Text Excerpt Showing Back-of-Book and Program-Chosen Index Terms (Cont.)



Padua, built only fifty years after Vesalius taught there. It is much as it was then. But the term "theater" is misleadingly spacious. For the room is a tiny, airless pit, oval in form and scarcely thirty feet across. Around the sides run shallow galleries in which one can barely stand. What must the atmosphere have been when these were packed with scores of sweating students, some of whom would surely faint or vomit, all jostling and craning to see down on the slab in their midst where the professor was dissecting the putrefying cadaver of some thief or beggar who would have been notably unsavory even when alive.

The success of Vesalius's course and of the book which embodied it was compounded of three elements: the authority of its information, the method of exposition, and the systematic approach. None of these was wholly novel, and Vesalius's essential contribution was the comprehensive skill with which he wove them into a corpus of anatomical practice rather than originality in any single detail or method. Vesalius himself made a great point of learning anatomy from bodies rather than books. And it is true that Greek humanism in antiquity and Christian teaching in the Middle Ages had created a powerful repugnance for opening the human body even in death. Nevertheless, Vesalius was far from having been the first anatomist to look inside his subject. Queen Elizabeth allowed the medical school at Cambridge three criminals a year. At the University of Bologna there was a standing rule in the fourteenth century that the medical students might procure cadavers for dissection, provided they did not belong to people who lived within thirty miles of Bologna. Indeed, the problem of the inadequate supply of bodies, like that of their rapid decomposition, was a handicap but not an absolute obstacle to research.

Fig. 6 "Art, Life, and Experiment" Text Excerpt Showing Back-of-Book and Program-Chosen Index Terms (Cont.)

Section 3
DOCUMENTATION

3.1 BPHRAS PROGRAM

Description: The BPHRAS parsing program is described in the 1968 and 1969 reports (Refs. 2 and 3), where there are sections on description, input and output format, theory, tables, and flow diagram. This year some changes were made to BPHRAS to produce output suitable for use by the INDEX program. Only the changes will be documented here. The new output for the INDEX program is similar to that described in the 1969 report (Ref. 3) added to BPHRAS for EXTRACT, but it contains in addition to the word and its position, the noun phrase of which the word is a part. All words of the text are included in the output which are nonfunction nonadverbial words and are a part of a noun phrase. These are considered as candidates for selection by the INDEX program.

New Output: Output of the index words is performed by a logical IOCS data definition module called LLEIPO which in turn uses IOCS mode IJGFOZZZ. Records of 175 bytes each are put out on disk in a block of 1758 bytes. Each file contains the index words for one file processed by BPHRAS. Each record contains information for 1 possible index word; the first 22 bytes are the ECBDIC representation of the word, the next 2 bytes give the page number on which the word was found (binary, right justified), the next byte gives the sentence number in which the word was found (within the page, starting with zero), and the last 150 bytes give the EBCDIC representation of the complete noun phrase containing the word. Both the word and phrase are truncated or padded with blanks where necessary.

After BPHRAS is run, the possible index words must be removed from the scratch disk; they are sorted into alphabetic order and stored on mangetic tape. The record format is not changed. The SORT control cards are given below.

SORT Control Cards:

```
// JOB SORT WORD-PHRASE FROM DISK
// ASSGN SYS005,X'192'                 SORT WORK
// ASSGN SYS005,X'191'
// ASSGN SYS001,X'180'                 SORT OUTPUT
// ASSGN SYS001,X'181'
// ASSGN SYS004,X'192'                 SORT INPUT-BPHRAS INDEX OUTPUT FILEA
// ASSGN SYS004,X'191'
// VOL SYS001,FILEO
// TPLAB 'INDEX WORD FILE  XXXXXXXXXXXXXXXXXXXX 68365 XXXXXX       DOS36C
                 0 ISLMSC'
// DLBL FILEA,'INDEX WORD AND PHRASE FILE',,SD
// EXTENT SYS004,094737,1,,20,980
// DLBL FILEW,'SORT WORK AREA',,DA
// EXTENT SYS005,094737,1,,1000,980
// EXEC DSORT
 SORT FIELDS=(1,22,A,26,100,A),FORMAT=BI,FILES=1,SIZE=3000
 RECORD TYPE=F,LENGTH=(175)
 INPFIL INPUT=D,BLKSIZE=(1750,X)
 OUTFIL OUTPUT=T,BLKSIZE=1750
 OPTION PRINT,LABEL=(U,S)
 END
```

Control Cards: The control cards needed for ouput onto the scratch disk are given below:

```
// DLBL LLEIPO, 'INDEX WORD AND PHRASE FILE',, SD
// EXTENT SYS004, 094737, 1,, 20, 980
```

Flow Diagram Changes: Only the last page of the BPHRAS flow diagram given in the 1969 report (Ref. 3) is different (Fig. 7); it has been both changed and augmented and will be given here by two new pages.

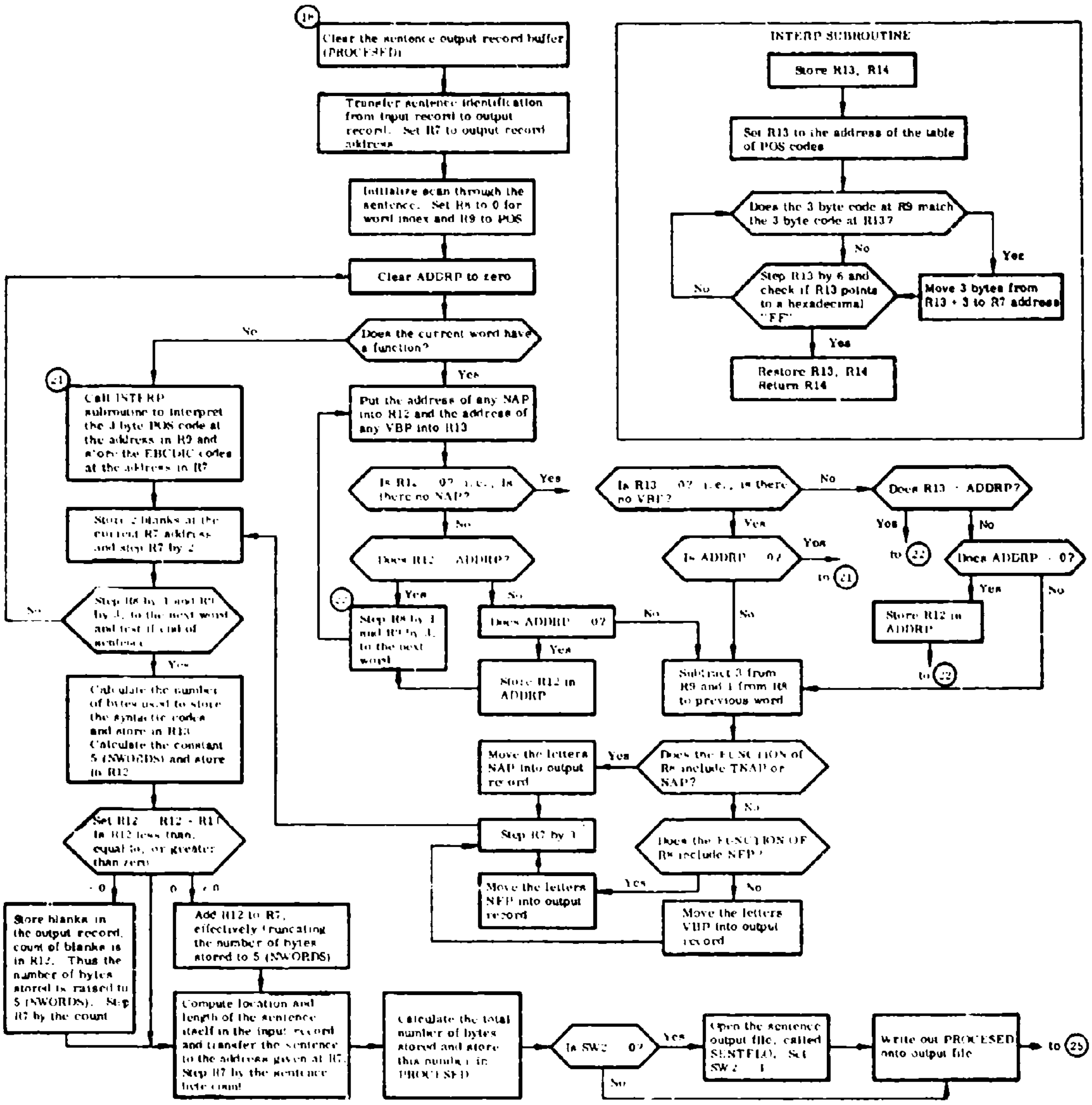

19
Clear the sentence output record buffer (PROCESED)
Transfer sentence identification from input record to output record. Set R7 to output record address
Initialize scan through the sentence. Set R8 to 0 for word index and R9 to POS
Clear ADDRP to zero
Does the current word have a function?
No
Yes
21
Call INTERP subroutine to interpret the 3 byte POS code at the address in R9 and store the EBCDIC codes at the address in R7
Store 2 blanks at the current R7 address and step R7 by 2
Step R8 by 1 and R9 by 3, to the next word and test if end of sentence
No
Yes
Calculate the number of bytes used to store the syntactic codes and store in R13. Calculate the constant 5 (NWORDS) and store in R12
Set R12 = R12 - R13. Is R12 less than, equal to, or greater than zero
&lt; 0
0
> 0
Store blanks in the output record, count of blanks is in R12. Thus the number of bytes stored is raised to 5 (NWORDS). Step R7 by the count
Add R12 to R7, effectively truncating the number of bytes stored to 5 (NWORDS)
Compute location and length of the sentence itself in the input record and transfer the sentence to the address given at R7. Step R7 by the sentence byte count
Calculate the total number of bytes stored and store this number in PROCESED
Is SW2 = 0?
Yes
No
Open the sentence output file, called SENTFLO. Set SW2 = 1
Write out PROCESED onto output file
to 25
Put the address of any NAP into R12 and the address of any VBP into R13
Is R12 = 0? i.e., is there no NAP?
Yes
No
Does R12 = ADDRP?
Yes
No
22
Step R8 by 1 and R9 by 3, to the next word
Does ADDRP = 0?
Yes
No
Store R12 in ADDRP
Is R13 = 0? i.e., is there no VBP?
No
Yes
Does R13 = ADDRP?
Yes
to 22
No
Is ADDRP = 0?
Yes
to 21
No
Does ADDRP = 0?
Yes
No
Store R12 in ADDRP
to 22
Subtract 3 from R9 and 1 from R8 to previous word
Does the FUNCTION of R8 include TNAP or NAP?
Yes
No
Move the letters NAP into output record
Step R7 by 3
Does the FUNCTION OF R8 include NFP?
Yes
No
Move the letters NFP into output record
Move the letters VBP into output record
INTERP SUBROUTINE
Store R13, R14
Set R13 to the address of the table of POS codes
Does the 3 byte code at R9 match the 3 byte code at R13?
No
Yes
Step R13 by 6 and check if R13 points to a hexadecimal "FF"
No
Yes
Move 3 bytes from R13 + 3 to R7 address
Restore R13, R14
Return R14

Fig. 7 BPHRAS Flow Diagram (Cont.)

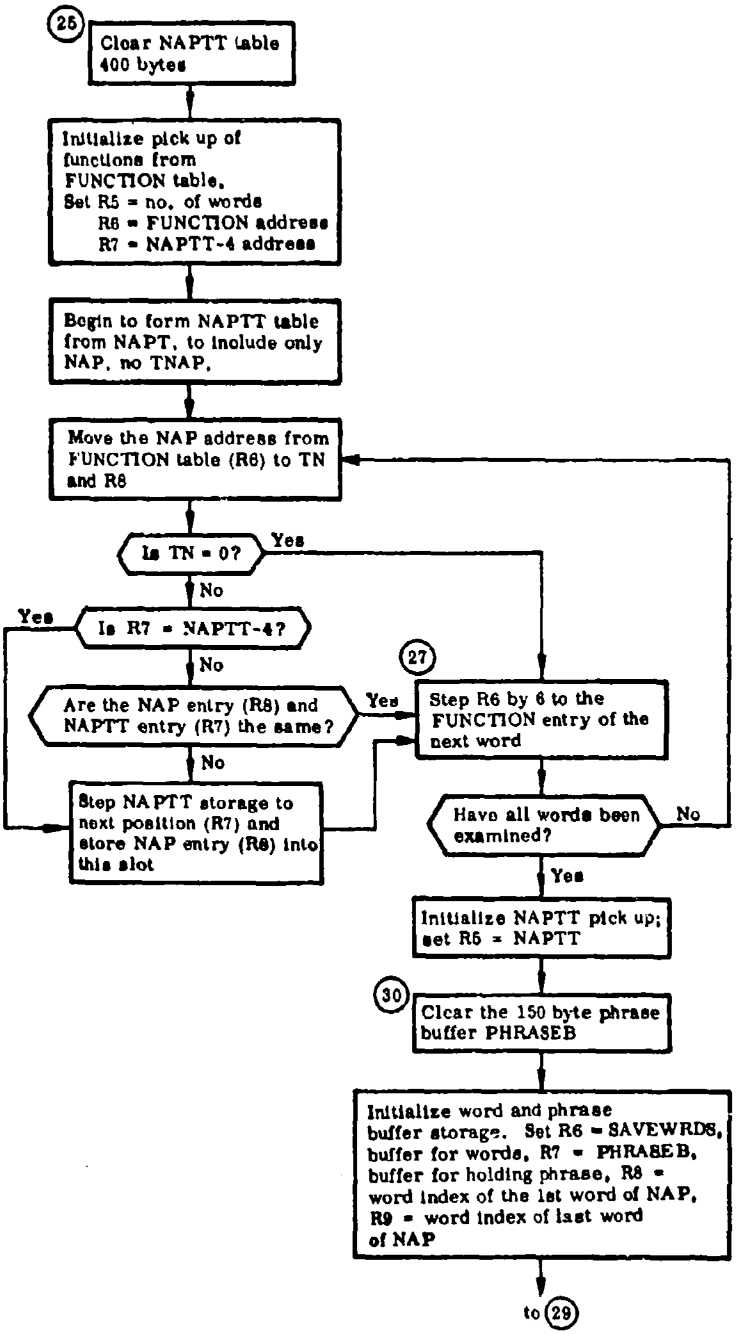

26
Clear NAPTT table 400 bytes
Initialize pick up of functions from FUNCTION table.
Set R5 = no. of words
R6 = FUNCTION address
R7 = NAPTT-4 address
Begin to form NAPTT table from NAPT, to include only NAP, no TNAP.
Move the NAP address from FUNCTION table (R6) to TN and R8
Is TN = 0?
Yes
No
Is R7 = NAPTT-4?
Yes
No
Are the NAP entry (R8) and NAPTT entry (R7) the same?
Yes
No
Step NAPTT storage to next position (R7) and store NAP entry (R8) into this slot
27
Step R6 by 6 to the FUNCTION entry of the next word
Have all words been examined?
No
Yes
Initialize NAPTT pick up; set R5 = NAPTT
30
Clear the 150 byte phrase buffer PHRASEB
Initialize word and phrase buffer storage. Set R6 = SAVEWRDS, buffer for words, R7 = PHRASEB, buffer for holding phrase, R8 = word index of the 1st word of NAP, R9 = word index of last word of NAP
to 29

Fig. 7 BPHRAS Flow Diagram (Cont.)

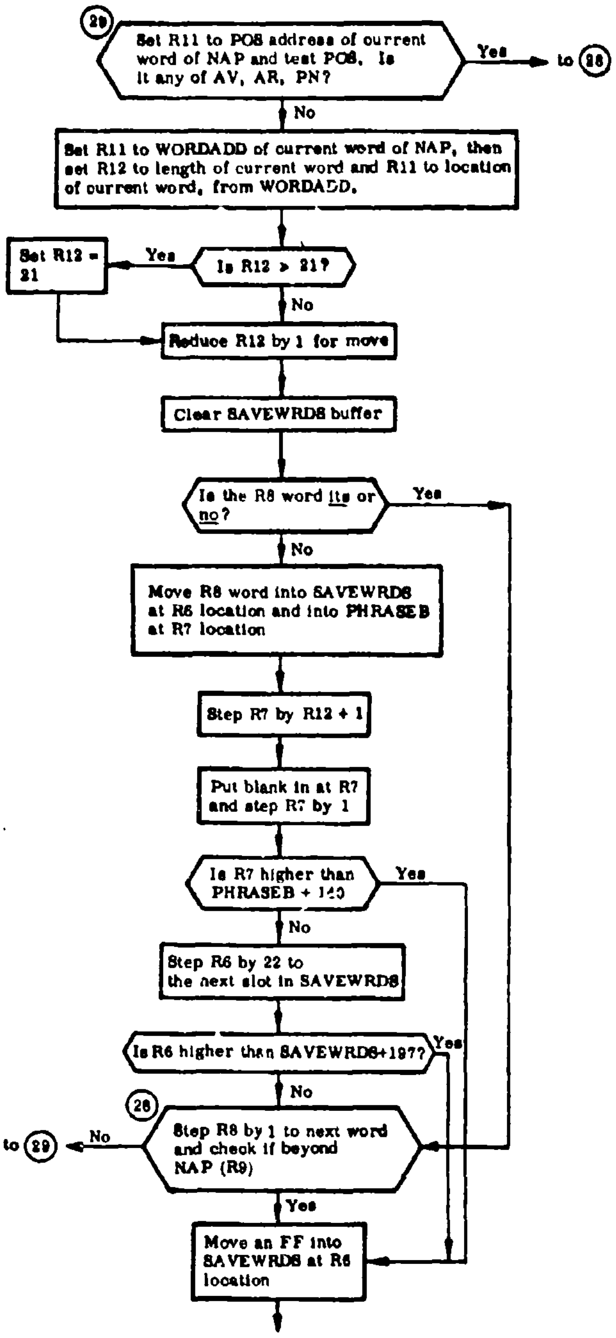
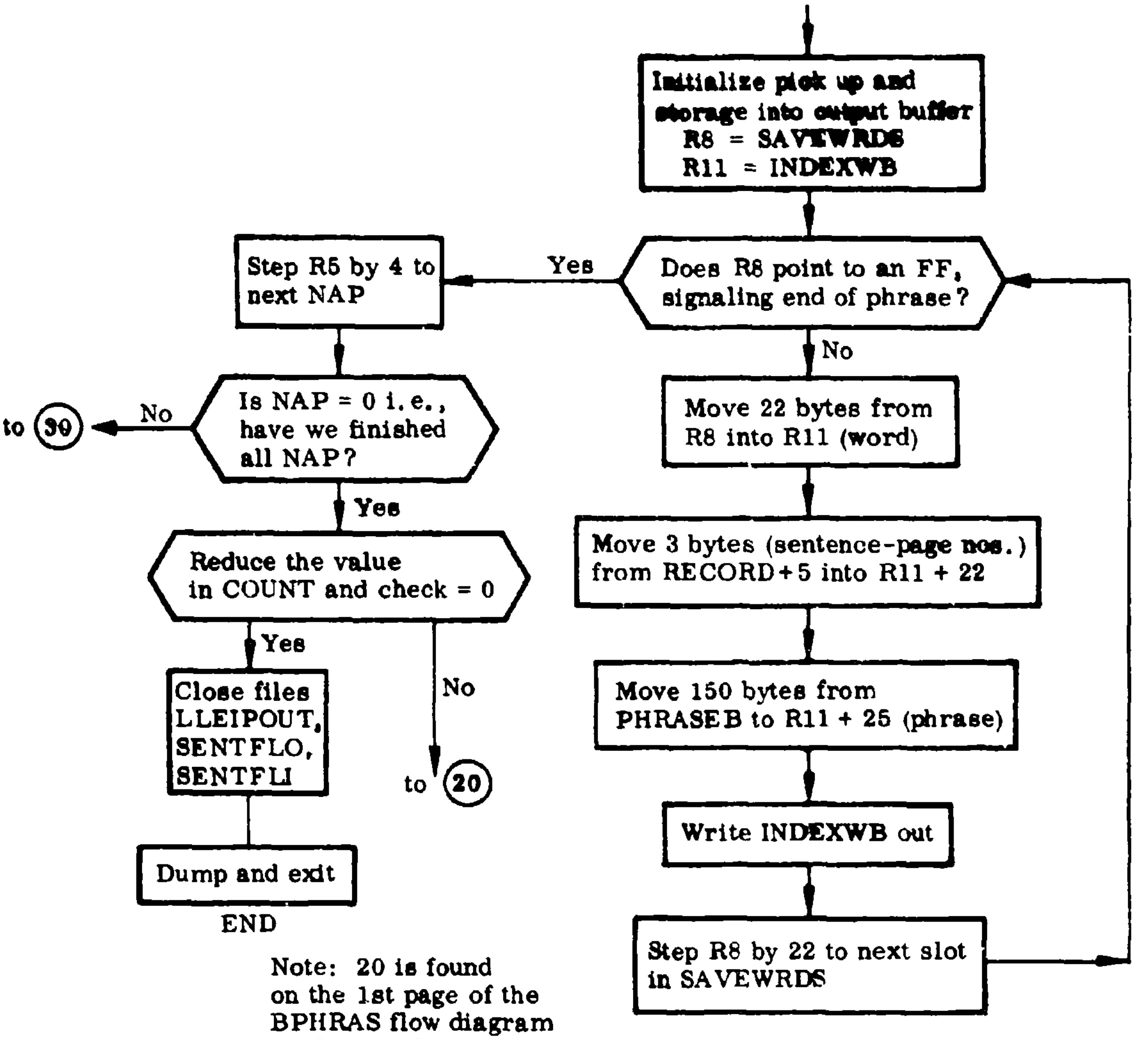

29
Set R11 to POS address of current word of NAP and test POS. Is it any of AV, AR, PN?
Yes
to 28
No
Set R11 to WORDADD of current word of NAP, then set R12 to length of current word and R11 to location of current word, from WORDADD.
Is R12 > 21?
Yes
Set R12 = 21
No
Reduce R12 by 1 for move
Clear SAVEWRDS buffer
Is the R8 word its or no?
Yes
No
Move R8 word into SAVEWRDS at R6 location and into PHRASEB at R7 location
Step R7 by R12 + 1
Put blank in at R7 and step R7 by 1
Is R7 higher than PHRASEB + 140
Yes
No
Step R6 by 22 to the next slot in SAVEWRDS
Is R6 higher than SAVEWRDS+197?
Yes
No
28
Step R8 by 1 to next word and check if beyond NAP (R9)
No
to 29
Yes
Move an FF into SAVEWRDS at R6 location

Fig. 7 BPHRAS Flow Diagram (Cont.)


Initialize pick up and storage into output buffer
R8 = SAVEWRDS
R11 = INDEXWB
Does R8 point to an FF, signaling end of phrase?
Yes
No
Step R5 by 4 to next NAP
Is NAP = 0 i.e., have we finished all NAP?
No
to 30
Yes
Reduce the value in COUNT and check = 0
Yes
No
to 20
Close files LLEIPOUT, SENTFLO, SENTFLI
Dump and exit
END
Move 22 bytes from R8 into R11 (word)
Move 3 bytes (sentence-page nos.) from RECORD+5 into R11 + 22
Move 150 bytes from PHRASEB to R11 + 25 (phrase)
Write INDEXWB out
Step R8 by 22 to next slot in SAVEWRDS

Note: 20 is found
on the 1st page of the
BPHRAS flow diagram
(Ref. 2) and is the point
at which another sentence
is read and processing begun

Fig. 7 BPHRAS Flow Diagram (Cont.)

3.2 DUPPHRAS PROGRAM

Description: The DUPPHRAS program examines a sentence file to select out, interpret, and print all contiguous sentences which have the same phrase structure as another adjacent sentence. In other words, it prints out all sentences with duplicate phrase structure. The program works on a sentence file which is produced by BPHRAS and then sorted into alphabetic order on the phrase-structure field and stored on magnetic tape. The SORT routine (control cards given in section 3.1) sorts on a fixed field of 100 bytes extending from the phrase structure field into the alphabetic field of the sentence. This does result in the rare loss of a sentence in a group. For example, in the unlikely arrangement below, the duplication of (1) and (3) would be missed. Usually there are more than two in a duplicate set and this problem does nct arise.

(1) NAP VBP PU KILROY WAS HERE.
(2) NAP VBP PU NAP PU THERE ARE....
(3) NAP BVP PU VERY FEW CAN DECIDE.

Input and Output: Input consists of a sorted sentence file on an unlabeled magnetic tape as is described in section 3.1. The IOCS file definition module SENTFL reads one record for each call, into the buffer specified in the calling sequence. Records are blocked 2000 records to the block, and are variable in length. The record format is described of BPHRAS output, section 5.3 of the 1968 report (Ref. 2).

Output is on-line on the printer. It is buffered, using an IOCS module called PRINT. Register 3 is used to transmit the buffer location to the program. The first line output for each sentence gives the sentence category in column 2 (I for indexible, N for nonindexible), the page number in columns 9−16, and the sentence number in columns 25−32. The second line is blank. Starting with line three, the codes showing the phrase structure of the sentence are printed. Two kinds of codes are used, phrase codes and word or punctuation codes, separated by blanks. These codes are listed in the BPHRAS documentation, section 5.3 of the 1968 report (Ref. 2). After the structure codes, the sentence itself is printed.

Control: There are two switches in the program which can be set by a REP card before running. A hexidecimal 47 FO stored at the address 28 FC will cause a skip to beginning or middle of page before printing instead of just spacing three lines. A 47 FA stored at 28B8 will cause a transfer around the check and deletion of print of sentences identical to one already printed.

Control Cards:

```
// JOB DUPPHRAS
// ASSGN SYS002, X'180'
// OPTION LINK, DUMP
        PHASE DUPPHRAS, ROOT
        INCLUDE
        PROGRAM DECK
/*
// EXEC LNKEDT
// EXEC
```

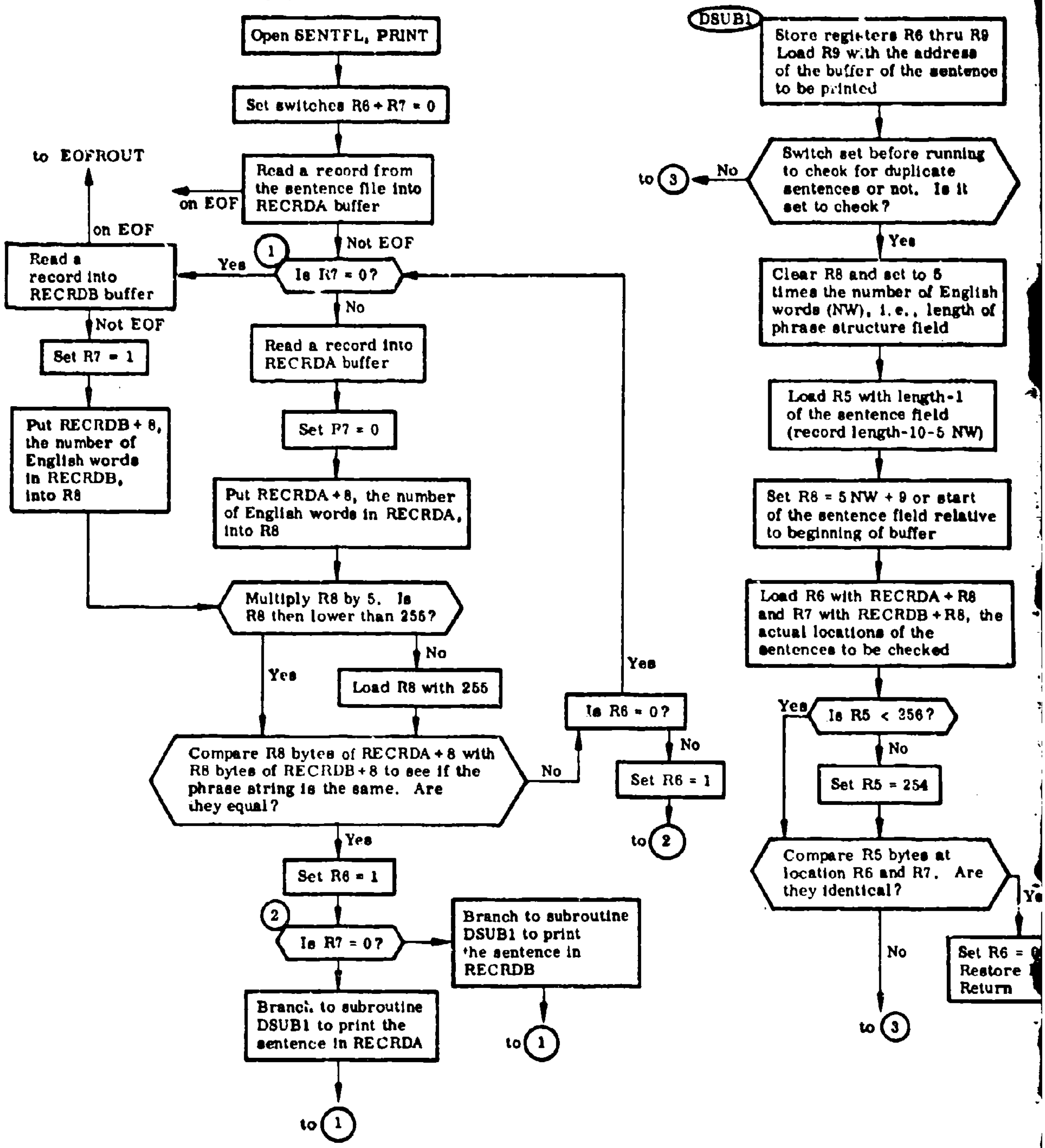

Open SENTFL, PRINT
Set switches R6 + R7 = 0
to EOFROUT
Read a record from the sentence file into RECRDA buffer
on EOF
on EOF
Not EOF
Read a record into RECRDB buffer
Yes
1
Is R7 = 0?
No
Not EOF
Set R7 = 1
Read a record into RECRDA buffer
Set P7 = 0
Put RECRDB + 8, the number of English words in RECRDB, into R8
Put RECRDA + 8, the number of English words in RECRDA, into R8
Multiply R8 by 5. Is R8 then lower than 256?
No
Yes
Load R8 with 255
Yes
Is R6 = 0?
No
Compare R8 bytes of RECRDA + 8 with R8 bytes of RECRDB + 8 to see if the phrase string is the same. Are they equal?
No
Set R6 = 1
to 2
Yes
Set R6 = 1
2
Is R7 = 0?
Branch to subroutine DSUB1 to print the sentence in RECRDB
Branch to subroutine DSUB1 to print the sentence in RECRDA
to 1
to 1
DSUB1
Store registers R6 thru R9 Load R9 with the address of the buffer of the sentence to be printed
Switch set before running to check for duplicate sentences or not. Is it set to check?
to 3
No
Yes
Clear R8 and set to 5 times the number of English words (NW), i.e., length of phrase structure field
Load R5 with length-1 of the sentence field (record length-10-5 NW)
Set R8 = 5 NW + 9 or start of the sentence field relative to beginning of buffer
Load R6 with RECRDA + R8 and R7 with RECRDB + R8, the actual locations of the sentences to be checked
Yes
Is R5 < 256?
No
Set R5 = 254
Compare R5 bytes at location R6 and R7. Are they identical?
Ye
No
Set R6 =
Restore
Return
to 3

Fig. 8 Flow Diagram for DUPPHRAS

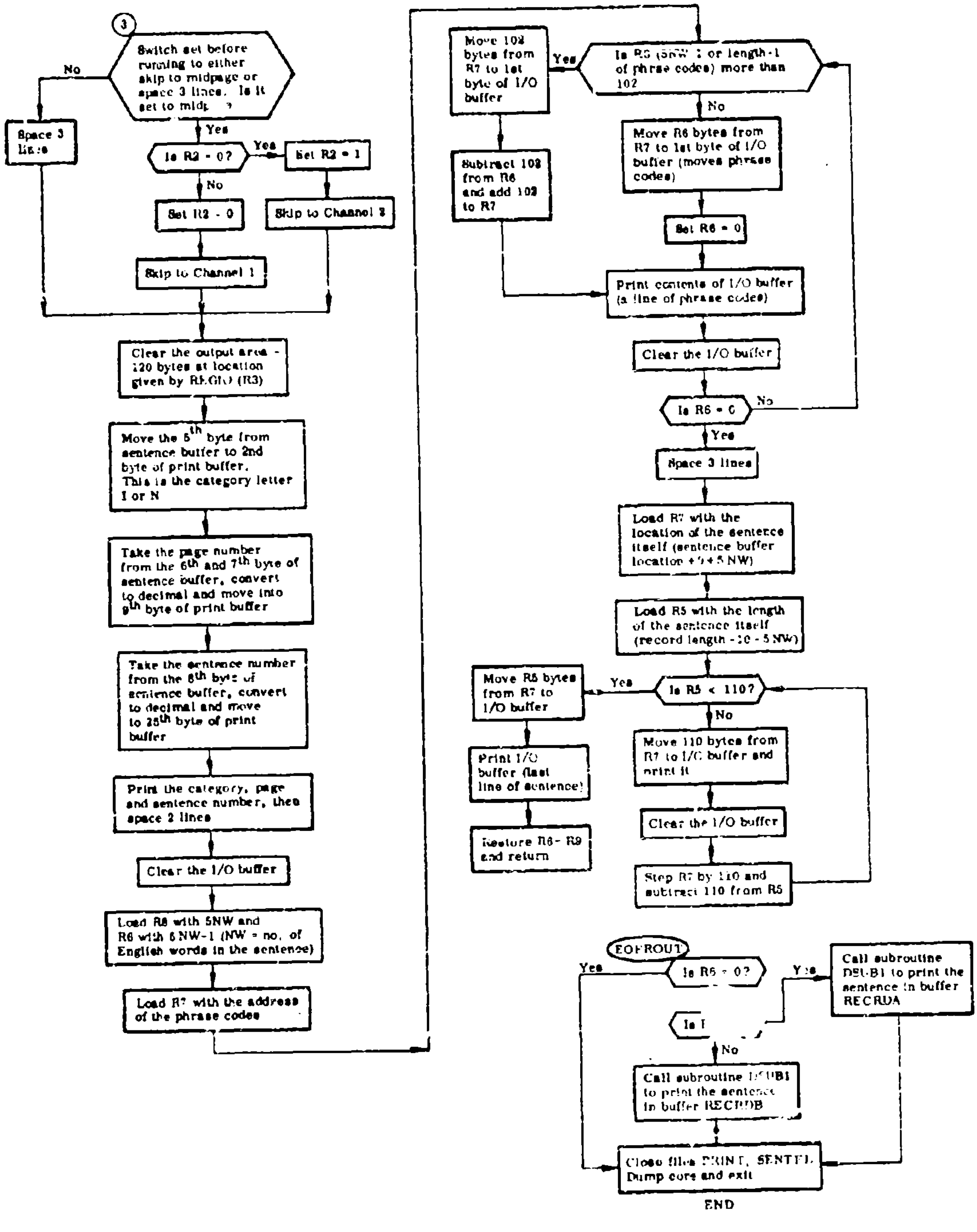

3
Switch set before running to either skip to midpage or space 3 lines. Is it set to midp ?
No
Space 3 lines
Yes
Is R2 - 0?
Yes
Set R2 = 1
No
Set R2 - 0
Skip to Channel 2
Skip to Channel 1
Clear the output area - 120 bytes at location given by REGIO (R3)
Move the 5th byte from sentence buffer to 2nd byte of print buffer. This is the category letter I or N
Take the page number from the 6th and 7th byte of sentence buffer, convert to decimal and move into 9th byte of print buffer
Take the sentence number from the 8th byte of sentence buffer, convert to decimal and move to 25th byte of print buffer
Print the category, page and sentence number, then space 2 lines
Clear the I/O buffer
Load R8 with 5NW and R6 with 5NW-1 (NW = no. of English words in the sentence)
Load R7 with the address of the phrase codes
Move 102 bytes from R7 to 1st byte of I/O buffer
Yes
Is R6 (5NW - 1 or length-1 of phrase codes) more than 102
No
Subtract 102 from R6 and add 102 to R7
Move R6 bytes from R7 to 1st byte of I/O buffer (moves phrase codes)
Set R6 = 0
Print contents of I/O buffer (a line of phrase codes)
Clear the I/O buffer
Is R6 = 0
No
Yes
Space 3 lines
Load R7 with the location of the sentence itself (sentence buffer location + 9 + 5 NW)
Load R5 with the length of the sentence itself (record length - 10 - 5 NW)
Move R5 bytes from R7 to I/O buffer
Yes
Is R5 < 110?
No
Print I/O buffer (last line of sentence)
Move 110 bytes from R7 to I/O buffer and print it
Restore R6 - R9 and return
Clear the I/O buffer
Step R7 by 110 and subtract 110 from R5
EOFROUT
Yes
Is R5 = 0?
Yes
Call subroutine DSUB1 to print the sentence in buffer RECRDA
Is I
No
Call subroutine DSUB1 to print the sentence in buffer RECRDB
Close files PRINT, SENTFL Dump core and exit
END

Fig. 8 Flow Diagram for DUPPHRAS (Cont.)

3.3 INDEX PROGRAM

Description: The INDEX program produces an index from text that has been previcusly processed by the parsing program BPHRAS. One of the outputs of BPHRAS is a list of all the words in the text which are found in a noun phrase and are not functional or adverbial in character. With each word is the complete noun phrase of which is is part. This list is sorted into alphabetic order and stored on tape for use by INDEX. (See section 3.1.) The INDEX program performs a frequency count on these words, and then calculates 2 possible frequency cutoff points, A and B. All words with a frequency of A or greater are then printed out, and underneath each word are all the noun phrases which contained that word, with their page and sentence number. This process is then repeated for all words with a frequency of B or greater. Theory and choice of cutoff points is discussed in section 2.3.

Input: Input comes in on magnetic tape on SYS008. See New Output in the BPHRAS documentation, section 3.1, for a complete description of the input. Input is accomplished through the IOCS definition module LLEINDI which in turn uses IOCS module IJFFZZWZ. The input tape is actually read twice, first for calculating frequency counts and then for selecting the actual index words and phrases chosen.

Output: The index words are output on the printer in alphabetic order. Under and indented from each index word are the noun phrases containing that word, also in alphabetic order according to the first word in the phrase. Within the grouping of phrases under an index word, a phrase is only printed once regardless of how many times it may have occurred in the text. If a phrase contains more than one index word, it will appear in the printout once under each index word. To the right of each phrase is a number representing the page and sentence number where the phrase appeared. If the phrase appeared more than once, these numbers are continued onto the next line or lines, interrupting the printing of the next phrase. The last three digits are the sentence number within the page. The first 9 digits give the number K+s, where K is a constant which indicates what text this came from and s is the sentence number. For example, 10025 indicates sentence 25 in the text whose constant is 10000.

Printing is accomplished with an IOCS definition module called PRINT which in turn uses IOCS module IJDFCPIZ. Printing is buffered. IOCS-program communication is via register 3.

<u>Control Cards</u>:

```
// JOB INDEX
// ASSGN SYS008, X '180 '
// ASSGN SYS008, X '181 '
// OPTION LINK, DUMP
        PHASE INDEX, ROOT
        INCLUDE
        Index binary deck
/*
// EXEC LNKEDT
// EXEC
```

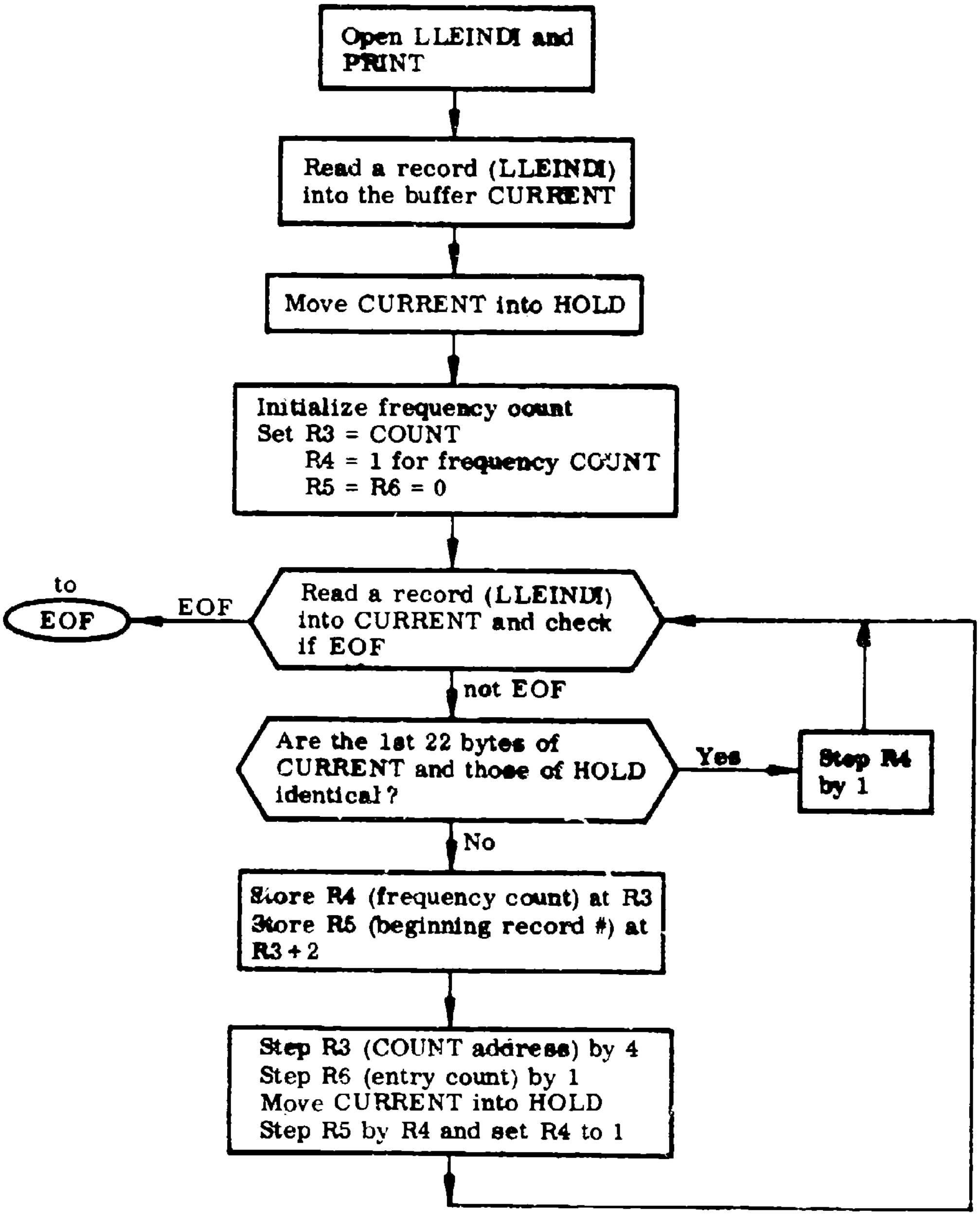

Open LLEINDI and PRINT
Read a record (LLEINDI) into the buffer CURRENT
Move CURRENT into HOLD
Initialize frequency count
Set R3 = COUNT
R4 = 1 for frequency COUNT
R5 = R6 = 0
to
EOF
EOF
Read a record (LLEINDI) into CURRENT and check if EOF
not EOF
Are the 1st 22 bytes of CURRENT and those of HOLD identical?
Yes
Step R4 by 1
No
Store R4 (frequency count) at R3
Store R5 (beginning record #) at R3 + 2
Step R3 (COUNT address) by 4
Step R6 (entry count) by 1
Move CURRENT into HOLD
Step R5 by R4 and set R4 to 1

Fig. 9 Flow Diagram for INDEX Program

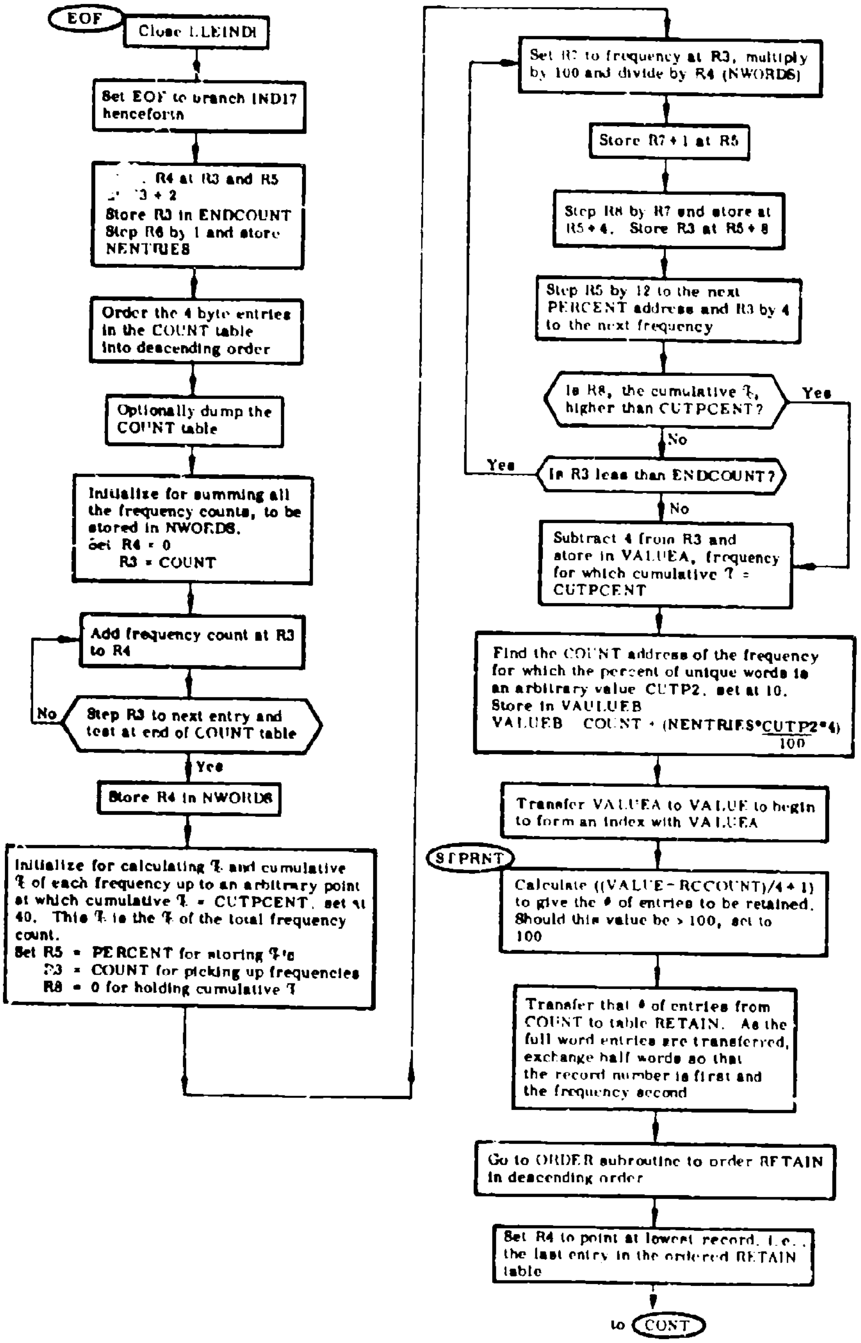

EOF
Close LLEINDI
Set EOF to branch IND17 henceforth
. R4 at R3 and R5 3 + 2 Store R3 in ENDCOUNT Step R6 by 1 and store NENTRIES
Order the 4 byte entries in the COUNT table into descending order
Optionally dump the COUNT table
Initialize for summing all the frequency counts, to be stored in NWORDS. Set R4 = 0 R3 = COUNT
Add frequency count at R3 to R4
No
Step R3 to next entry and test at end of COUNT table
Yes
Store R4 in NWORDS
Initialize for calculating % and cumulative % of each frequency up to an arbitrary point at which cumulative % = CUTPCENT, set at 40. This % is the % of the total frequency count.
Set R5 = PERCENT for storing %'s
R3 = COUNT for picking up frequencies
R8 = 0 for holding cumulative %
Set R7 to frequency at R3, multiply by 100 and divide by R4 (NWORDS)
Store R7 + 1 at R5
Step R8 by R7 and store at R5 + 4. Store R3 at R5 + 8
Step R5 by 12 to the next PERCENT address and R3 by 4 to the next frequency
Is R8, the cumulative %, higher than CUTPCENT?
Yes
No
Yes
Is R3 less than ENDCOUNT?
No
Subtract 4 from R3 and store in VALUEA, frequency for which cumulative % = CUTPCENT
Find the COUNT address of the frequency for which the percent of unique words is an arbitrary value CUTP2, set at 10. Store in VAULUEB
VALUEB COUNT + (NENTRIES*CUTP2*4)/100
Transfer VALUEA to VALUE to begin to form an index with VALUEA
STPRNT
Calculate ((VALUE - RCCOUNT)/4 + 1) to give the # of entries to be retained. Should this value be > 100, set to 100
Transfer that # of entries from COUNT to table RETAIN. As the full word entries are transferred, exchange half words so that the record number is first and the frequency second
Go to ORDER subroutine to order RETAIN in descending order
Set R4 to point at lowest record, i.e., the last entry in the ordered RETAIN table
to CONT

Fig. 9 Flow Diagram for INDEX Program (Cont.)

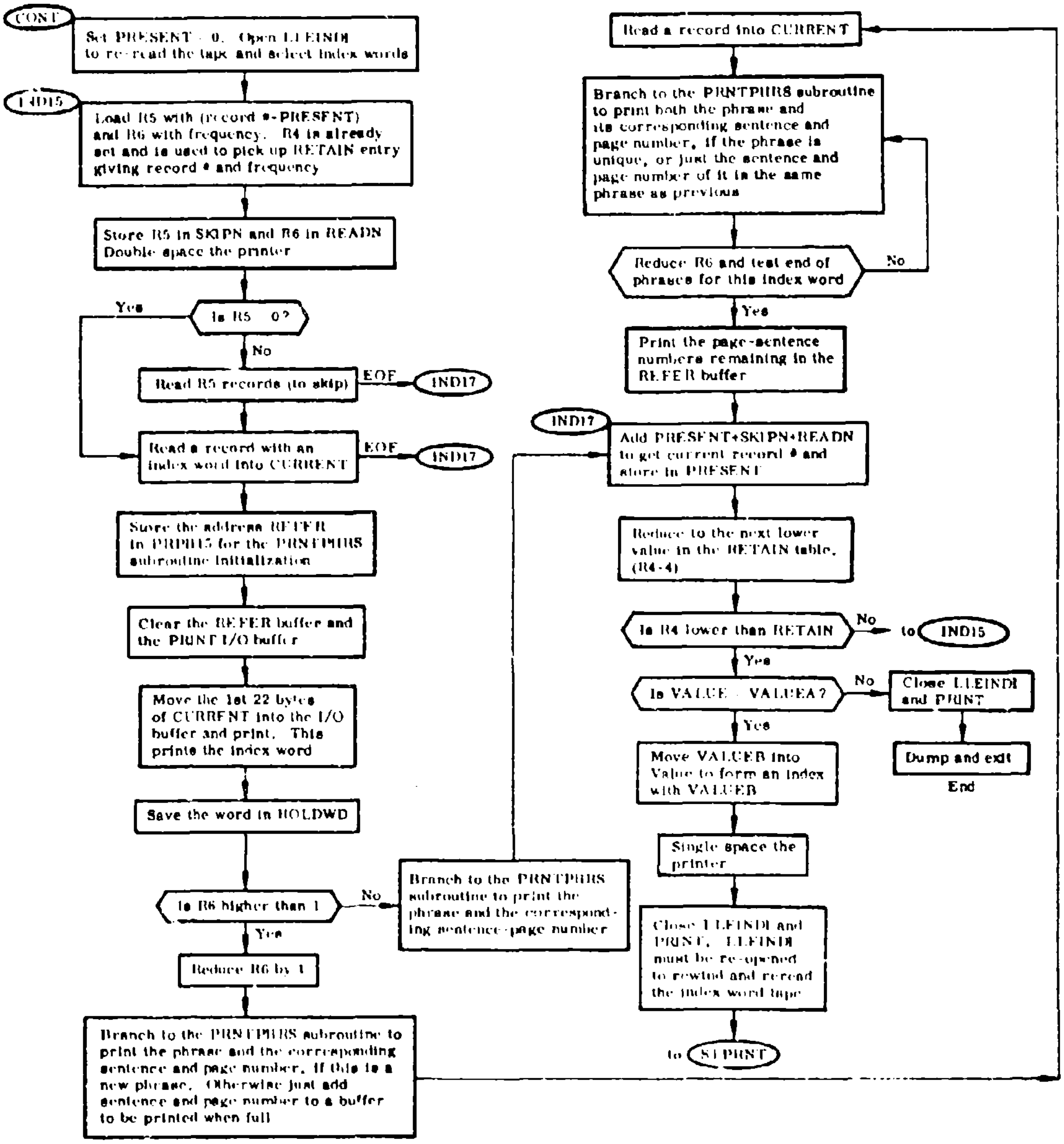

CONT
Set PRESENT = 0. Open LLEINDI to re-read the tape and select index words
IND15
Load R5 with (record #-PRESENT) and R6 with frequency. R4 is already set and is used to pick up RETAIN entry giving record # and frequency
Store R5 in SKIPN and R6 in READN Double space the printer
Is R5 = 0?
Yes
No
Read R5 records (to skip)
EOF
IND17
Read a record with an index word into CURRENT
EOF
IND17
Store the address REFER in PRPH15 for the PRNTPHRS subroutine initialization
Clear the REFER buffer and the PRNT I/O buffer
Move the 1st 22 bytes of CURRENT into the I/O buffer and print. This prints the index word
Save the word in HOLDWD
Is R6 higher than 1
No
Yes
Branch to the PRNTPHRS subroutine to print the phrase and the corresponding sentence-page number
Reduce R6 by 1
Branch to the PRNTPHRS subroutine to print the phrase and the corresponding sentence and page number, if this is a new phrase. Otherwise just add sentence and page number to a buffer to be printed when full
Read a record into CURRENT
Branch to the PRNTPHRS subroutine to print both the phrase and its corresponding sentence and page number, if the phrase is unique, or just the sentence and page number of it is the same phrase as previous
Reduce R6 and test end of phrases for this index word
No
Yes
Print the page-sentence numbers remaining in the REFER buffer
IND17
Add PRESENT+SKIPN+READN to get current record # and store in PRESENT
Reduce to the next lower value in the RETAIN table, (R4-4)
Is R4 lower than RETAIN
No
to IND15
Yes
Is VALUE = VALUEA?
No
Close LLEINDI and PRINT
Dump and exit
End
Yes
Move VALUEB into Value to form an index with VALUEB
Single space the printer
Close LLEINDI and PRINT. LLEINDI must be re-opened to rewind and reread the index word tape
to STPRNT

Fig. 9 Flow Diagram for INDEX Program (Cont.)

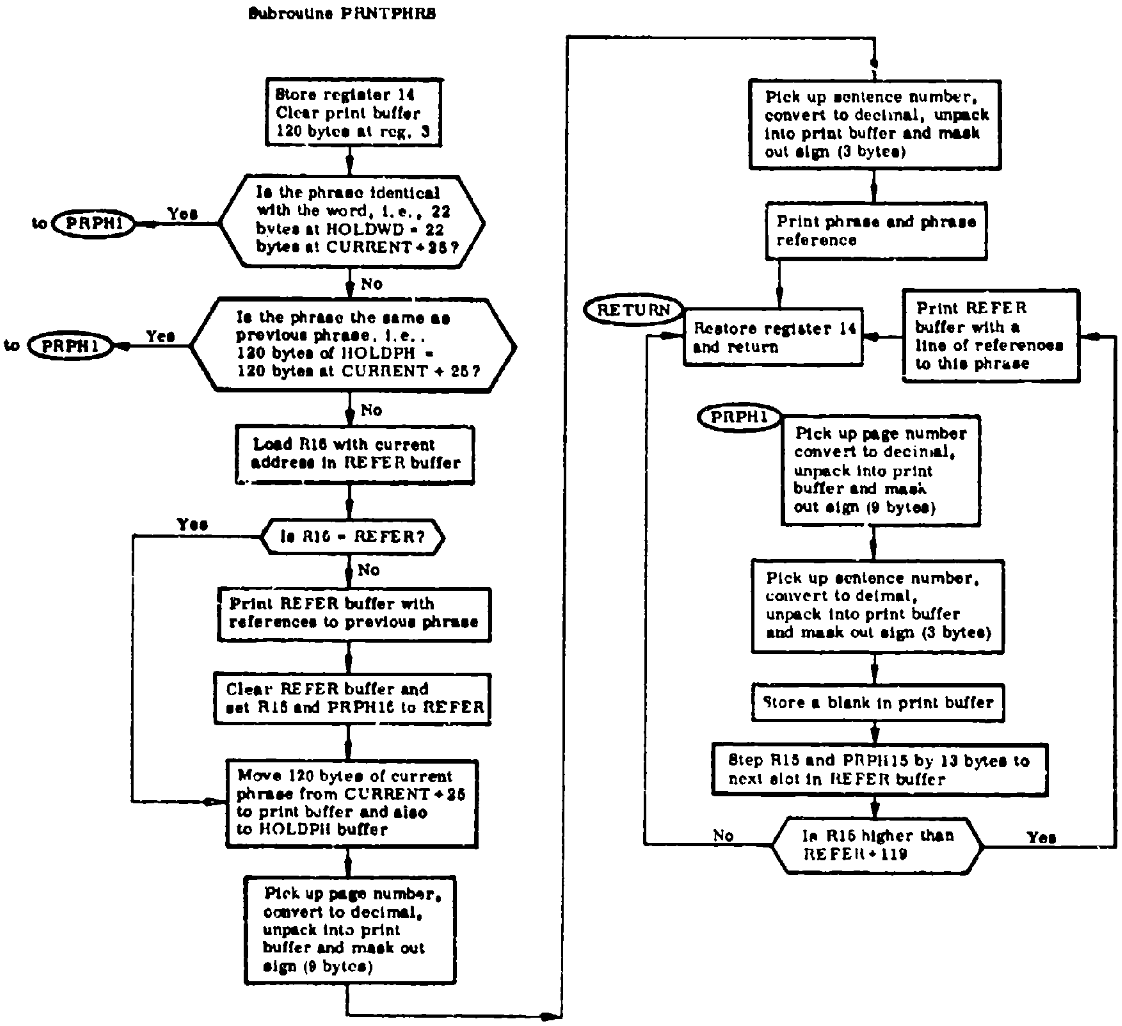
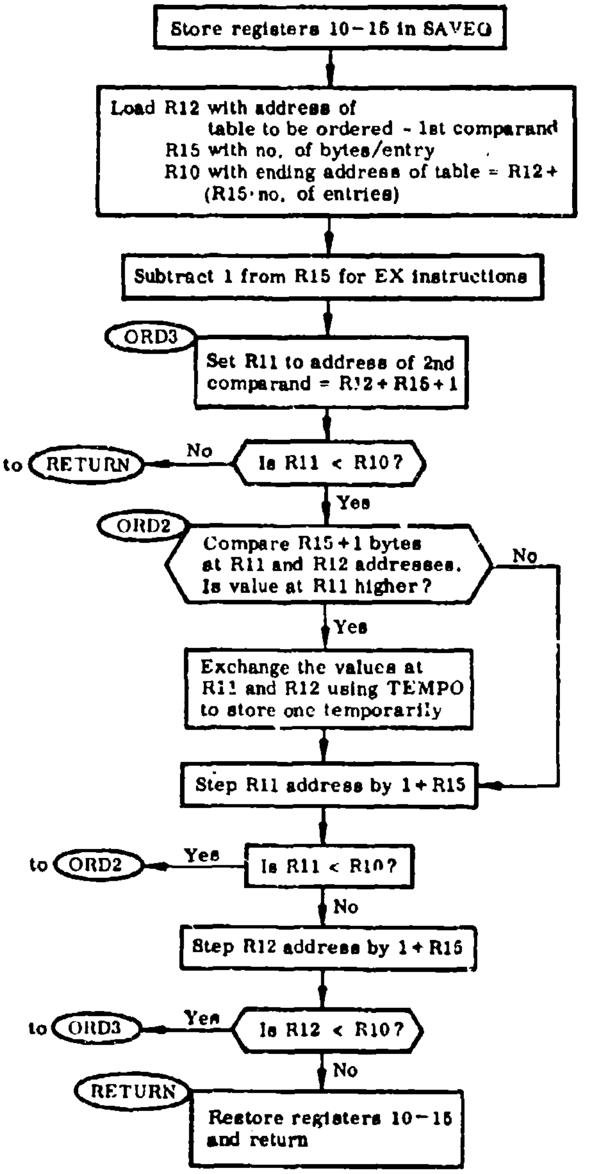

Subroutine PRNTPHRS
Store register 14 Clear print buffer 120 bytes at reg. 3
Is the phrase identical with the word, i.e., 22 bytes at HOLDWD = 22 bytes at CURRENT+25?
Yes
to PRPH1
No
Is the phrase the same as previous phrase, i.e., 120 bytes of HOLDPH = 120 bytes at CURRENT + 25?
Yes
to PRPH1
No
Load R15 with current address in REFER buffer
Is R15 = REFER?
Yes
No
Print REFER buffer with references to previous phrase
Clear REFER buffer and set R15 and PRPH15 to REFER
Move 120 bytes of current phrase from CURRENT+25 to print buffer and also to HOLDPH buffer
Pick up page number, convert to decimal, unpack into print buffer and mask out sign (9 bytes)
Pick up sentence number, convert to decimal, unpack into print buffer and mask out sign (3 bytes)
Print phrase and phrase reference
RETURN
Restore register 14 and return
Print REFER buffer with a line of references to this phrase
PRPH1
Pick up page number convert to decimal, unpack into print buffer and mask out sign (9 bytes)
Pick up sentence number, convert to delmal, unpack into print buffer and mask out sign (3 bytes)
Store a blank in print buffer
Step R15 and PRPH15 by 13 bytes to next slot in REFER buffer
No
Is R15 higher than REFER+119
Yes

Fig. 9 Flow Diagram for INDEX Program (Cont.)

Subroutine ORDER

Calling Routine: BAL 14, ORDER
4 byte address of table
2 byte parameter = no. of entries
2 byte parameter = no. bytes/
entry max = 20



Store registers 10-15 in SAVEQ
Load R12 with address of table to be ordered - 1st comparand
R15 with no. of bytes/entry
R10 with ending address of table = R12+ (R15·no. of entries)
Subtract 1 from R15 for EX instructions
ORD3
Set R11 to address of 2nd comparand = R12+R15+1
to RETURN
No
Is R11 < R10?
Yes
ORD2
Compare R15+1 bytes at R11 and R12 addresses. Is value at R11 higher?
No
Yes
Exchange the values at R11 and R12 using TEMPO to store one temporarily
Step R11 address by 1+R15
to ORD2
Yes
Is R11 < R10?
No
Step R12 address by 1+R15
to ORD3
Yes
Is R12 < R10?
No
RETURN
Restore registers 10-15 and return

Fig. 9 Flow Diagram for INDEX Program (Cont.)

3.4 FREQUENCY-SYNTAX INDEXING PROCEDURE

To obtain an index of a text using the INDEX program, the following steps are necessary:

(1) If the text already exists as a file with parts-of-speech assigned to each word, this step can be skipped. Otherwise, the SENDIC program must be run to form a sentence file for each text. SENDIC is described in the 1967 report (Ref. 1). To run, SENDIC cards containing the input sentences (see RSENTR for format, which is free-form), preceded if desired by a page card, must be read in through the card reader. The sentences with the assigned parts of speech will be output on tape.

(2) Run the BPHRAS program to obtain the possible index words and their noun phrases, which will be stored on disk. (See section 3.1 and also Ref. 2 and Ref. 3 for BPHRAS writeup.) The BPHRAS input and output are given below, with the relevant items starred.

* Tape unit 180 – The sentence file produced by SENDIC, containing sentences with the parts-of-speech assigned to words.

Tape unit 181 – A scratch tape for output of a sentence file containing sentences with a phrase structure assigned.

* Disk unit 191 or 192 – A scratch disk for output of the possible index words and their noun phrases.

(3) Sort the words and phrases from the scratch disk onto tape, as described in section 3.1 of this report. The input and output here will be as follows:

Disk unit 191 or 192 – The scratch disk with possible index words and noun phrases put out by BPHRAS.

Tape unit 180 or 181 – The above words and phrases sorted into alphabetic order of the words.

(4) Run INDEX to select and print the words and phrases of the index. INDEX is described in section 3.3 of this report. The tape output of Step (3) is input to the program; the index words and phrases come out on the printer.

Section 4
REFERENCES FOR PART I

1. Annual Report: Automatic Indexing and Abstracting, M-21-67-1, Lockheed Palo Alto Research Laboratory, Mar 1967

2. Annual Report: Automatic Informative Abstracting and Extracting, M-21-68-1, Lockheed Palo Alto Research Laboratory, Mar 1968

3. Annual Report: Automatic Informative Abstracting and Extracting, M-21-69-1, Lockheed Palo Alto Research Laboratory, Mar 1969

4. L. L. Earl, "Automatic Determination of Parts of Speech of English Words," Mechanical Translation, Vol. 10, Nos. 3 and 4, Sep and Dec 1967

5. L. L. Earl, Annual Report: Automatic Indexing and Abstracting, Part I, M-21-66-1, Lockheed Missiles & Space Co., Mar 1966

6. H. P. Luhn, "The Automatic Creation of Literature Abstracts," IBM J. of Research and Development, Vol. 2, No. 2, 1958, pp. 159 – 165

7. (1) A. Doak Barnett, Communist China & Asia, Challenge to American Policy, published for the Council on Foreign Relations, Harper, New York, 1960

   (2) Neil W. Chamberlain, The Firm: Micro-Economic Planning and Action, McGraw-Hill, New York, 1962

   (3) Charles Coulston Gillispie, The Edge of Objectivity: an Essay in the History of Scientific Ideas, Princeton University Press, Princeton, N.J., 1960

   (4) Walter James Greenleaf, Occupations: A Basic Course for Counselors, Government Printing Office, Washington, D. C., 1954

   (5) John G. Gurley and Edward S. Shaw, Money In a Theory of Finance, Brookings Institution, Washington, D.C., 1960

   (6) Robert L. Heilbroner, The Making of Economic Society, Prentice-Hall, Englewood Cliffs, N. J., 1962

(7) Mark S. Massel, Competition and Monopoly, Legal and Economic Issues, Brookings Institution, Washington, D. C., 1962

(8) Robert M. Palter, Whitehead's Philosophy of Science, University of Chicago Press, 1960

(9) S. M. Siegel, The Plant Cell Wall – A Topical Study of Architecture, Dynamics, Compartive Chemistry, and Technology in a Biological System (International Series of Monographs of Pure and Applied Biology, Plant Physiology Division, Vol. 2), Pergamon, N. Y., 1962

8. R. K. Summit, "DIALOG: An Operational On-Line Reference Retrieval System," Proceedings ACM National Meeting, 1967

Part II

OBSERVATIONS ON COMPUTER-DETECTABLE SEMANTIC STRUCTURES

Section 5

INTRODUCTION

One of the major problems in natural-language processing is the problem of multiple meaning, that phenomenon in which a single word or word group shows itself capable of assuming a variety of semantic meanings. This paper discusses a set of English words, each one of which is able to convey its various meansings by the use of easily recognizable and adjacent syntactic structures such as, say, prepositional phrases. Indeed it is more accurate to say that among this set of words semantic meaning is generally inseparable from the word plus its associated syntactic structure.

The computer-detectability of these words lies in the fact that the associated syntactic structures which pinpoint a given word's semantic meaning almost always consist of a small group of prepositions or other function words which, by virtue of their small number and their adjacency to the word whose meaning they specify, are readily detectabie by computer algorithm.

This set of words whose semantics is syntax-linked is itself a subset of another set of words. Words in this larger set are called "government" words. The word "government" derives from the ability of these words to impose constraints – usually syntactic – on words surrounding them. The original intent of the word-government project was syntactic in nature; however, as compilation proceeded it was realized that in a substantial number of cases there was a linkage with semantic meaning. Therefore, before the subject of syntax-related semantics can be discussed, it will be necessary to review word-government.

Section 6

WORD GOVERNMENT

Human language can be viewed as a vehicle for describing relationships. When a dictionary identifies a word as possessing a certain part of speech, a relationship is in fact being defined. Thus denoting hit as a transitive verb means that it requires a direct object. The relationships denoted by parts of speech are binary in nature (I hit the ball). While words possessing word-government also define binary relationships, in general such words define relationships between three, four, and, occasionally, five elements. In more conventional grammatical parlance, government words are words capable of governing two or more objects.

Word-government can be seen in the following sentence: I believe in his ability to pass the bar exam.

Believe governs in his ability. But note that ability also governs the infinitive to pass. Hence there are two instances of word-government in this sentence which illustrate how a network of government-linked words may be established.

Believe governs other words and phrases besides the preposition in. Some of these words and phrases are: I believe him, I believe that he will come, I believe he will come, I believe what he said, I believe in what he said. The relationship between believe and the elements which it governs can be conveniently shown in the following arrangement:

believe	vt	S
	vt	(that) + clause
	vt	what + clause

vi	in S
vi	in what + clause
vt	S/ (to be) S

where vt denotes a transitive verb, vi denotes an intransitive verb, S denotes a substantive, and the parentheses indicate that the use of that and to be is optional.

In the same way the government relationships of ability and comparison are shown below:

ability	n	inf
comparison	n	of S/ with S
	n	of S/ to S
	n	between S/ and S
	n	of S/ and S

where n denotes a noun, S a substantive, and inf denotes an infinitive.

Following the Ramo-Wooldridge nomenclature [1], the governing words (believe, ability, comparison, etc.) will be called primaries. The elements governed by the primaries (S, that + clause, in S, etc.) will be called secondaries or secondary patterns. Thus the phenomenon of word-government deals with two sets of English words – the set of primaries and the set of secondaries – and the relationships between the two sets. Primaries are nouns, verbs, adjectives, and, occasionally, adverbs. Secondaries are usually prepositional phrases, clauses, a particular form of the verb (e.g., the infinitive form), and certain case forms of nouns (especially important in Russian).

The primary-secondary relationship can be readily seen in these arrangements. If the secondary pattern contains more than one element, then the elements of the pattern are separated by slashes.

When secondaries occur on one line, they are compatible with each other; that is, they may all occur together in a given sentence. Secondaries on different lines, as with *comparison*, are incompatible; one or the other, but not both, may be associated with their primary in a given sentence.

The following examples show that primaries may be nouns, verbs, adjectives, and adverbs.† The designation, vip, of *amuse* denotes an intransitive verb whose pattern (*at S*) occurs only in the passive voice.

Primary	Part of Speech	Secondary Pattern
depend	vi	on S/ for S
	vi	on S/ inf
	vi	on S/ to be S
	vi	on CJ* + clause
translate	vt	S/ from S/ into S
	vt	S/ from S/ to S
associate	vt	S/ with S
	vt	S/ to S
	vt	Px/ with S**
	vt	Px/ with S/ in S
	vi	with S
transformation	n	of S/ from S/ into S
	n	of S/ from S/ to S
immune	aj	to S
	aj	from S
	aj	against S
careful	aj	of S
	aj	with S
	aj	inf
	aj	CJ + clause
enroute	av	from S/ to S
	av	for S
amuse	vip	at S

*CJ denotes a conjunctive, a term used by Hornby [2] to represent the set of interrogative adverbs and pronouns (how, what, when, where, who, whom, whose, why).

**Px denotes a reflexive pronoun (himself, herself, etc.).

†It is often necessary to use inflectional forms of the primaries given in the word government table. For example, although "at S" is given as an entry for "amuse," the inflectional form "amused" must be used for this, e.g., "amused at the book."

The original intent of the word-government project was purely syntactic in nature – to compile a list of English primaries and their associated secondary syntactic patterns. For example, in the original compilation the government patterns for elevation and incidence appeared as

elevation	n	of S/ to S
incidence	n	of S/ in S
	n	of S/ among S
	n	of S/ on S
	n	of S/ upon S

However, as compilation proceeded it was realized that in a substantial number of cases the secondary patterns tended to be linked with semantic meaning. This phenomenon became so pronounced that it was finally decided to recompile the word-government listing and attempt, where possible, to associate secondary patterns with semantic meaning. In the recompilation and in the examples which follow, semantic meanings are denoted by arabic numerals placed to the left of the secondary. Thus the patterns for elevation and incidence now appear as follows:

elevation	1	n	of S
	2	n	of S/ to S
incidence	1	n	of S/ in S
	1	n	of S/ among S
	2	n	of S/ on S
	2	n	of S/ upon S

where elevation (1) is synonymous with height (except in the architectural sense)

The building has an elevation of 1,000 feet.

But elevation (2) describes the action of raising

The newspapers announced his elevation to the peerage.

Incidence (1) is similar to occurrence.

They reported a high incidence of suicide in Sweden.

Incidence (2) describes the act, manner, or fact of falling upon or influencing.

> The incidence of light on a reflecting surface can be measured by precise instruments.

The compilation of the English word-government listing is sufficiently advanced (about 70% complete) to allow an estimate of the final size of the dictionary. There will be approximately 8000 primaries; an average primary will have five secondary patterns associated with it. Thus there will be approximately 40,000 distinct entries (an entry being a primary plus one secondary pattern) in the dictionary.

Section 7

SEMANTIC STRUCTURES

A semantic structure is a word-government complex (i.e., a primary word plus its associated secondary patterns) in which the secondary patterns are used to convey semantic distinctions in the primary governing word. The previous examples, elevation and incidence, are semantic structures.

7.1 PRIMARY AND SECONDARY ELEMENTS

Nouns and verbs make up well over 90% of the primary words functioning in semantic structures. One-syllable verbs in particular possess a very wide "semantic spectrum" but here, too, we will see that the secondary patterns are extensively used to convey meaning.

In the set of semantic structures, adjectives may occasionally function as the primary word

intent	1	aj	S (e.g., an intent young man...)
	2	ajp	on S (e.g., he was intent on murder)

where aj and ajp denote attributive and predicative forms of the adjective. The incidence of such adjectives, however, is quite small compared to the number of nouns and verbs.

It is the secondary pattern, distinguishing as it does between the primary's various meanings, which plays the key role in a semantic structure. Common secondary patterns which serve in a semantics-distinguishing role are single prepositions, combinations of single prepositions, certain forms of the verb (e.g., the infinitive), and the relative pronoun that. In addition to these common secondary elements, many other types of secondary patterns occur in semantic structures. They are discussed in the next section.

7.2 EXAMPLES OF SEMANTIC STRUCTURES

The various types of semantic structures will now be discussed in terms of their characterizing secondary patterns. The government dictionary has only been compiled through P so none of the examples given fall in the latter part of the alphabet.

Before beginning, however, it would perhaps be helpful to illustrate a word-government structure which is not a semantic structure. All transitive verbs have, by definition, the following pattern

verb vt S

where, as before, S represents a substantive. Now when a preposition is compatible with this basic transitive pattern, the new pattern may only define more precisely the original meaning.

thank	vt	S
	vt	S/ for S

He thanked his friend.
He thanked his friend for the favor.

buy	vt	S
	vt	S/ for S

He bought a book.
He bought a book for three dollars.

The additional information supplied by the prepositions does not cause any semantic alteration in the basic meaning of the two verbs, thank and buy.

The pattern S/ prep S: In the following examples, semantic alteration does take place with the addition of the indicated prepositions.

index	1	vt	S
	2	vt	S/ under S

(1) He indexed the book (i.e., he prepared an index of the book)

(2) He indexed the book under 'medicine' (i.e., the book was catalogued under 'medicine')

administer	1	vt	S
	2	vt	S/ to S

(1) She needs someone to administer her affairs (i.e., to manage or direct her affairs)

(2) He administered the pill to the patient (i.e., gave or dispensed it)

head	1	vt	S
	2	vtx	S/ for S
	2	vtx	S/ toward S
	2	vtx	S/ into S
			⋮

(1) He heads the company. (i.e., He manages the company)

(2) He headed the ship for (toward, into) open water (i.e., He pointed the ship for open water)

The dots in the patter of head indicate that the word-government dictionary contains additonal patterns which have been omitted as not being relevant to the present discussion.

7.2.1 Single Prepositions

The above pattern, S/ prep S, is very common; however, single prepositions may also function alone as semantic discriminators.

extract	1	n	from S
	2	n	of S

Extract (1) is: a passage which has been extracted from some text. Extract (2) is: that which has been obtained, usually by crushing or pressing.

keep	1	vi	from S
	1	vi	S/ from S
	2	vi	at S
	3	vi	to S
			⋮

Keep (1) means: to prevent, keep (2) means: to perserve or to maintain at, keep (3) means: to continue moving in a specified direction.

The use of single prepositions as semantic discriminators is especially evident in one-syllable verbs as we will see below.

7.2.2 Infinitive

The infinitive (which will henceforth be denoted in secondary patterns as "to-inf") is a frequently used semantic discriminator in secondary patterns.

habituate	1	vt	S
	2	vt	S/ to S
	2	vt	S/ to-inf

Habituate (1) means: to frequent a place, habituate (2) means: to acclimate.

indisposed	1	ajp	0
	2	ajp	to S
	2	ajp	to-inf

where ajp denotes a predicative adjective and 0 denotes lack of a pattern. Indisposed (1) means: to be ill (i.e., He is indisposed). Indisposed (2) means: disinclined (He is indisposed to do any work today).

intend	1	vt	to-inf
		⋮	
	2	vt	S/for S

Where intend (1) means: to have (a purpose) in mind (I intend to finish the job); intend (2) means: to be destined for (The house is intended for his son).

entitle	1	vt	S/ S
	2	vtx	S/ to S
	2	vtx	S/ to-inf

Entitle (1) means: to name a book, a play... (Somerset Maugham entitled his first book Liza of Lambeth). Entitle (2) means: to have earned, to have coming (His accomplishments entitle him to a vacation. His accomplishments entitle him to take a vacation).

Another type of infinitive, the so-called "bare infinitive" [2] may also play a role as a semantic discriminator.

dare	1	vtx	to-inf
	1	vtx	bare-inf
	2	vtx	S
	3	vtx	S/ to-inf

Dare (1) means: to be brave enough to do something (He dared to sail around the world). With the bare infinitive the "to" is dropped (He dared sail around the world). Dare (2) means: to face, to take the risk of (He dared the rapids). Dare (3) means: to challenge (He dared his enemy to attack).

Reflexive Pronouns

Reflexive pronouns – denoted by PX – may serve as semantic discriminators (in the pattern below D denotes an adverb).

deport	1	vt	S/ to S
	2	vtx	PX/ D
	2	vtx	PX/ with S

Deport (1) means: to send someone out of the country. Deport (2) means: to behave (He deported himself badly).

disengage	1	vt	S/ from S
	2	vtx	PX/ from S

Disengage (1) means: to separate something from something else. Disengage (2) means: to disassociate oneself from something or someone.

establish	1	vt	S/ in S
	1	vt	S/ at S
			⋮
	2	vtx	PX/ as S

Establish (1) means: to set something up, to found something. Establish (2) means: to prove oneself to be something.

7.2.4 That + Clause

That followed by a clause is a frequently used semantic discriminator.

admission	1	n	of S/ into S
	1	n	of S/ to S
	2	n	of S/ (that) + clause
	2	n	by S/ (that) + clause

As before, the parentheses indicate that the use of that is optional and, in fact, if that is not used, then the clause assumes the role of semantic discriminator. Admission (1) means: being admitted into or to something. Admission (2) is: a confession.

Other examples of that + clause are listed below.

grant	1	vt	S/ S
	1	vt	S/ to S
	2	vt	(that) + clause
			⋮

Where grant (1) is: to bestow and grant (2) is equivalent to: given that (I grant that he is intelligent...)

gather	1	vt	S/ from S/ to S
	2	vt	(together) S/ into S
	3	vt	S
	4	vtx	from S/ (that) + clause

Here gather (4) means: to infer (I gather that you will be at the party).

insinuate	1	vtx	S/ into S
	1	vtx	PX/ into S
	2	vt	(that) + clause
	2	vt	to S/ that + clause

where insinuate (1) is: to gain admission or someone's confidence by stealth.

Insinuate (2) means: to suggest something in an unpleasant manner.

7.2.5 Gerunds and Participles

Gerunds and present participles, denoted by g and PR, respectively, occasionally find use as semantic discriminators though their use is not common.

defer	1	vt	G/ until S
	2	vi	to S

where defer (1) means: to postpone (He deferred writing the letter until...) defer (2)

hear	1	vt	S/ PR
	2	vt	S/ about S/ from S

where hear (1) means to perceive sound with the ears (He heard the car coming).

In the following example both gerund and present participle occur in the same set of patterns.

bargain	1	vi	with S/ for S
			⋮
	2	vi	on S/ PR
	2	vi	on G
			⋮

where bargain (1) is: to haggle, and bargain (2) means: to anticipate (We didn't bargain on John coming when he did).

7.2.6 Adjectives

Adjectives – denoted by A – are not common as semantic discriminators though like gerunds and participles they are used on occasion.

fade	1	vtx	S/ A
			⋮
	2	vi	into S
			⋮

where fade (1) means: to cause to grow pale (The sun faded the shirt white) and fade (2) means: to grow pale (Night faded into day).

extraction	1	n	of S/ from S
	2	n	A-

where the dash in the second pattern indicates that the adjective is to be used attributively. Extraction (1) means: removal, while extraction (2) usually signifies origin (He was of French extraction).

feel	1	vi	of S
	1	vt	S/ with S
			⋮
	4	vi	A

where feel (1) means: to touch, and feel (4) describes someone's physical or mental state (I feel good).

7.2.7 Conjunctions

Sometimes the conjunction and serves a minor role in semantic discrimination though this seems to be in a subsidiary role to the prepositions between and among.

discriminate	1	vt	between S/ and S
			⋮
	2	vi	against S
			⋮

where discriminate (1) means: to make, see a difference and discriminate (2) means: to treat differently.

7.2.8 Complete Examples

Many of the preceding examples were incomplete having been edited to emphasize usage of the particular type of secondary pattern being discussed. The following examples give complete sets of semantics-discriminating secondary patterns and are typical of the way that the patterns are used. In these examples the semantic meaning is listed to the right of the pattern.

deliver	1	vt	S/ to S	..........to take something someplace
	2	vt	S/ from S	.....to rescue, save
	2	vt	S/ out of S	
	3	vtx	PX/ of S	.......to give forth in words
	3	vt	S	
	4	vt	NM	.............to help in childbirth
	5	vt	/up S/ to S	...to surrender, give up
	5	vt	/over S/ to S	

NM in the fourth pattern denotes an animate noun. It should be noted that the category of animate noun is not, like the categories of other secondary elements, a syntactic category. This is the only nonsyntactic category used so far in compiling the government dictionary. In deliver (5) the slash which precedes the prepositions denotes a "floating" adverbial particle which may either follow or precede the substantive.

We backed up the car.
We backed the car up.

He took off his coat.
He took his coat off.

Verbs which allow this type of structure are actually two-word transitive verbs subject to the transformation

$$N_1\ V_1\ Av\ N_2 \rightleftarrows N_1\ V_t\ N_2\ Av$$

and must be distinguished from intransitive verbs with prepositional phrases. These two-word verbs are a common occurrence in the word-government dictionary.

determination	1	n	of S/ by S	............determining or being determined
	2	n	of S/ in S	........calculation, finding out
	2	n	of S/ from S	
	3	n	of S/ to-inf	...resolution, firmness of purpose
	3	n	of S/ that + clause	

The preposition of can follow most – though not all – nouns. When of is used to describe the possessive relationship then the noun which is the prepositional object of of is normally transformed into the possessive and placed in front of the lead noun (e.g., "John's determination to succeed" rather than "the determination of John to succeed"). In the word-government dictionary the possessive of* is placed in the secondary pattern for the sake of consistency – because all the other prepositions have been placed there, following the primary governing noun which they modify.

*Like most prepositions, of can describe a variety of relationships of which the possessive is but one. This is discussed below.

engage	1	vt	S/ as S/ to-inf	hire, employ as
	2	vt	S/ for S	to undertake
	2	vt	S/ to-inf	
	3	vi	in S/ with S	to participate, take part in
	4	vip	to S	to promise to marry
	5	vi	in S	to be occupied with
	5	vi	with S	
	6	vip	by S	to have the attention drawn by
	7	vt	S	to attack
	8	vt	S	to fit into, to fit together (esp. of machinery)

Note that engage (7) and engage (8) coincide; there does not appear to be any syntactic means of distinguishing between the two meanings.

head	1	vt	S	to manage something, to be at the top of
	2	vt	S/ for S	to move in the direction indicated
	2	vt	S/ toward S	
	2	vt	S/ into S	
	2	vi	for S	
	2	vi	toward S	
	2	vi	into S	
	3	vt	/off S	to get in front of so as to turn aside

The phenomenon of syntax-related semantics becomes especially evident among the one-syllable verbs. These words are especially important because it is difficult if not impossible to converse or write colloquially without them. The patterns for fall are typical of the many patterns this class of veib may govern.

fall	vi	1	from S/ to S/ onto S
	vi	1	from S/ to S/ into S
	vi	1	from S/ on S
	vi	1	from S/ upon S
	vi	1	out of S/ to S
	vi	1	out of S/ into S
	vi	1	out of S/ on S
	vi	1	out of S/ upon S
	vi	2	down S
	vi	2	among S
	vi	2	around S
	vi	2	against S
	vi	2	toward S
	vi	2	through S
	vi	2	off S
	vi	7	back/ before S
	vi	3	back on S
	vi	3	back upon S
	vi	4,2	behind S
	vi	5	behind on S
	vi	6,2	down on S
	vi	8	for S
	vi	9	in with S
	vi	10,2	short of S
	vi	2	in
	vi	11	off
	vi	12	on S/ to-inf
	vi	12	to S/ to-inf
	vi	13,2	through
	vi	14,2	under S

The numbers which follow the part-of-speech symbol denote the following semantic meanings:

(1) To drop from a higher to a lower place

(2) To drop against, among, etc. an object or group of objects

(3) To have recourse to

(4) To lose ground

(5) To become in arrears of

(6) To collapse, fail

(7) To retreat

(8) To be attracted, to love

(9) To agree, to meet

(10) To fail in

(11) To decline

(12) To become responsible

(13) To miscarry

(14) To be classified

But even the fourteen meanings found for fall seem miniscule compared with the seventy meanings of get most of which are syntax-related. Get is listed in Table 1. Go has 108 meanings, the largest number found so far.

7.2.9 The Semantics of Prepositions

We have now seen how the secondary government patterns can act as determiners of their primary word's semantic meanings. But what about the semantic content of the secondary elements themselves? Do these elements, like other English words, partake of a variety of meanings? And if they do how does this affect the foregoing discussion of semantic structures?

The fact is that prepositions, by far the most common elements comprising the secondary patterns, are indeed polysemic. A preposition not only connects an antecedent with its object, but also establishes a specific relation between the two. For a

Table 1

FORMS OF "GET"

1	vt	S/from S/for S	32	vi	NP down
1	vt	S/S/for S	33	vt	/down S
2	vt	S/from S	34	vi	in/through S
3	vt	S/at S/for S	34, 35	vi	in/by S
4	vt	S/to-inf	34	vi	into S/through S
5	vt	S/to S	34, 35	vi	into S/by S
6	vt	S/from S/to S/by S	36	vi	in/DT
6	vt	S/from S/to S/on S	37	vt	/in S
7	vt	S	38	vi	in on S/with S
8	vt	S	39	vi	into S
8	vt	what + cl	40	vi	off/with S
9	vt	S/in S/for S	40	vi	off/PR
9	vt	S/at S/for S	40	vtx	S off/with S
10	v	P/– –/it	41	vtx	S off
10	vtx	S	42	vt	/off S/to S
10	vtx	what + cl	43	vi	(up) on S
11	vt	S/S/for S	44	vt	S on
11	vt	S/S/to S	45	vi	on about S
12	vi	to S/*at S	45	vi	on with S
12	vi	in S/*at S	46	vi	on to S/by S
12	vi	into S/*at S	47	vt	/out S
12	vi	in on S/*at S	48	vi	out S/on S
13	vt	S/through S	49	vi	out of S
13	vt	S/in S	50	vi	out/
13	vt	S/into S	51	vtx	S over (with)
13	vt	S/off (of) S	52	vi	over S
13	vt	S/onto S	53	vi	through S
13	vt	S/up S	50	vt	S/through S
13	vt	S/A	50	vt	S/by S
13	vt	S/on S	55	vi	to S
14	vt	S/with S	56	vi	together (with) S/for S
15	vi	about S	57	vi	together (with) S/on S
15, 16	vi	around S	58	vt	S together
17	vt	/across S	59	vt	/up S/at S
18	vt	/across S/to S	60	vi	up/from S
19	vi	ahead/of S	61	vtx	PX up
20	vi	ahead/in S	62	vi	to-inf
21	vi	along/with S	63	vi	A
22	vi	at S	64	vt	S/A
23	vi	away/with S	64	vt	S/PS
24	vt	S away/from S	65	vt	S/to-inf
25	vi	away with S	65	vt	S/PS
25	vi	by with S	66	vt	P/(D)
26	vt	/back S/from S	66	vt	P/in S
27	vt	S/(back) to S	67	vprp	S
28	vi	(back) to S	68	vprp	to-inf
29	vi	by S/with S	69	vtx	PR
30	vi	down/from S	69	vtx	to PR
30	vi	off S/at S	70	vtx	S/for S
31	vi	down to S			

given preposition, this relation may vary with the result that a single preposition describes a variety of different relations between the antecedent and object. In an interesting discussion of the subject, Newman [3] examined the common prepositions beginning with "A." As an example of the large number of relations in which a common preposition may participate, at, in Newman's study, is shown as able to describe fifteen different relationships between antecedent and object!

The different functions of the preposition are sometimes distinguished in the government listing by simply creating another secondary pattern.

Thus, hire could be listed as

hire verb S/ for S/ at S

However, for S may be used in two different ways in this pattern.

We hired him for farm work at three dollars an hour.

We hired him for three dollars an hour.

This double usage would be better reflected by creating another secondary pattern for hire

hire verb S/ for S/ at S
S/ for S

Newman distinguishes different prepositional relations by superscripts. Thus the above pattern might be written as

hire verb S/ for^{2} S/ at^{5} S
S/ for^{7} S

where for and at (hypothetically) possess a spectrum of relationships and in this spectrum for^2 = relationship of purpose

for^7 = relationship of price

at^5 = relationship of price

This type of nomenclature would permit very precise definitions of the secondary patterns. Patterns which at present appear identical will be seen as distinct. Thus

hire	vt	S/ for S
admire	vt	S/ for S

might become

hire	vt	S/ for^7 S
admire	vt	S/ for^3 S

(We hired him for farm work. We admire him for his bravery.)

The ideal solution would be to denote each preposition in the government dictionary with a number (as Newman has done) which specifies the nature of the relationship between object and antecedent. Unfortunately, no such complete description of English prepositions exists. The compilation begun at the patent office was never completed [4]. Therefore, while the importance of prepositional semantics in precisely defining the secondary patterns is recognized, this feature has not yet been added to the government dictionary.

Let us suppose, nevertheless, that prepositional relationships have been defined (at least insofar as it is possible to do so) and that these prepositional meanings have been incorporated into the dictionary. Thus for hire

hire	vt	S/ for^2 S/ at^5 S
	vt	S/ for^7 S
		⋮

Such a structure will provide a very precise definition of word-government and will also provide a data-base especially amenable to computer sorting. From such a computer examination, many aspects of word-government which thus far have remained hidden may reveal themselves. For example, it is the author's intuitive opinion, after compiling several thousand primaries and their associated government patterns, that only certain prepositional relationships participate in the government phenomenon. For example, out of all of at's fifteen meaning-relationships, perhaps only five or six occur in government patterns.

7.2.10 Function Words

The specifically named words in the secondary patterns (e.g., the prepositions as opposed to the substantives) belong to the set of so-called "function words" whose frequency is so high that they are assumed to be without significance. Function words have received short shrift from information scientists who tend to regard them as a noninformation bearing matrix in which is embedded the "real stuff" of language. This view, while popular, is by no means universal. R. Pagés [5] has stated "...the linking terms of a message are not simply a sort of conjunctive tissue surrounding the 'telegram' ... but form a considerable part of its very substance."

More specifically, Wallace [6] has shown that differences in the frequency rankings of function words and other high frequency words were sufficient to allow the use of rank order of common words as a satisfactory means of classifying a given corpus of text as belonging to either the field of psychology or the field of computer science.

Meadow [7] has described a hypothetical processor which, by selecting from a piece of text the ten more frequent words and then comparing them with a previously prepared list of common words for various natural languages, is able to determine the particular natural language in which the text is written. The comparison list for the various languages is restricted to certain types of function words – articles, pronouns, prepositions, and conjunctions.

Bratley and Dakin, at the University of Edinburgh, are working on a limited dictionary for syntactic analysis, many of whose entries consist of certain verbs whose government properties are used to detect semantic ambiguities [8].

The phenomenon here called "word-government" is coming to be recognized as vital to an understanding of natural language processing. Thus Pagés has said [5] of verbal government: "... statistically the verb is a more polyvalent predicate than other parts of speech (especially adjectives) and extends its influence to a greater number of arguments. It is perhaps on account of its polyvalence that it generally carr'es ... common indications concerning the whole of the sentence: time, including time-aspects, character of assertion, modality of assertion, eventual negation ...."

Tesniere [9] has formulated a theory concerning the importance of words in which words are considered more significant in structurally more subordinate passages. And this polyvalence, this structural subordination, are in fact generally manifested by use of certain kinds of function words, especially prepositions.

Section 8

CONCLUSION

Two principal ideas have been demonstrated. First, there exists a set of English words each of which utilizes certain associated syntactic structures to convey its various semantic meanings; because the associated syntactic structures are signaled by easily recognizable and adjacent function words, these complete structures may readily be recognized by computer algorithm. Second, the phenomenon of semantic structures discussed here demonstrates that function words also have a semantic component, that they are used by their particular governing word to help the governing word convey its various semantic meanings. In fact, the use of function words as semantic discriminators raises a most interesting question: are function words so common precisely because they are needed by the language to discriminate semantic meaning?

As far as practical applications are concerned, the word-government dictionary should prove valuable to grammarians, both traditional and transformational, for it is a rich source of information not generally found in dictionaries. For example, the government dictionary lists those verbs which, though considered transitive, cannot be transformed into the passive voice; the dictionary also denotes whether a given verb's syntactic pattern occurs in the active or passive voice (e.g., amaze + at occurs only in the passive voice).

It was, of course, always hoped that the word-government dictionary would prove useful in natural language processing. The most immediate intended application at Lockheed Information Sciences Laboratory is to provide high-level automatic indexes. In such an application, the governing word will be linked, by means of its secondary patterns, to textual keywords and phrases thus establishing a network of relationships between the governing word and the keywords occurring in the government pattern. Beyond this, the word-government dictionary appears to have unique applications in the automatic rewording of sentences and as a detector and resolver of certain kinds of sentence ambiguity.

Section 9
REFERENCES FOR PART II

[1] Thompson Ramo Wooldridge, Inc., Machine-Translation Studies of Semantic Techniques, Contract AF 30(602)-2036, 22 Feb 1961

[2] A. S. Hornby, A Guide to Patterns and Usage in English, Oxford University Press, 1953

[3] Simon M. Newman, Analysis of Prepositionals for Interrelational Concepts, Preliminary Study, Patent Office Research and Development Reports, No. 16, 15 July 1959

[4] Simon M. Newman, Personal Communication, 2 May 1967

[5] R. Pagés, Relational Aspects of Conceptualization in Message Analysis, Information Storage and Retrieval, December 1967

[6] E. M. Wallace, Rank Order Patterns of Common Words as Discriminators of Subject Content in Scientific and Technical Prose, SP-1505, Santa Monica, Calif., System Development Corporation, April 1964

[7] Charles T. Meadow, The Analysis of Information Systems, John Wiley and Sons, Inc., 1967

[8] P. Bratley and D. J. Dakin, A Limited Dictionary for Syntactic Analysis, in Machine Intelligence 2, edited by Ella Dale and Donald Michie, American Elsevier Publishing Co., New York, 1968

[9] Lucien Tesniere, Elements de Syntaxe Structurale, 1966, Paris, Klincksieck

Unclassified
Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY *(Corporate author)*
Lockheed Palo Alto Research Laboratory
Lockheed Missiles & Space Company
Palo Alto, California 94304

2a. REPORT SECURITY CLASSIFICATION
Unclassified

2b. GROUP
N/A

3. REPORT TITLE
Annual Report: Automatic Informative Abstracting and Extracting

4. DESCRIPTIVE NOTES *(Type of report and inclusive dates)*
Annual Progress Report, March 1969 – March 1970

5. AUTHOR(S) *(First name, middle initial, last name)*
Lois L. Earl
Harold R. Robison

6. REPORT DATE
March 1970

7a. TOTAL NO. OF PAGES
84

7b. NO. OF REFS
17

8a. CONTRACT OR GRANT NO.
Nonr 4440(00)

b. PROJECT NO.

c.

d.

9a. ORIGINATOR'S REPORT NUMBER(S)
M-21-70-1

9b. OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)*

10. DISTRIBUTION STATEMENT
This document is subject to special export controls and each transmittal to foreign governments or foreign nationals may be made only with prior approval of the Office of Naval Research, Code 437.

11. SUPPLEMENTARY NOTES

12. SPONSORING MILITARY ACTIVITY
Department of the Navy
Office of Naval Research
Washington, D.C. 20360

13. ABSTRACT

Part I of this report documents several experiments in automatic indexing. Nine chapters, each from a different technical book were used as the text copies for all the experiments. In the first experiment, an attempt was made to construct a sentence dictionary of syntactic sentence types, for distinguishing extract-worthy sentences, but it proved unrewarding. Alogrithms developed to combine syntactic and statistical criteria in the choice of extract sentences and index phrases proved more rewarding. Extract sentences and index noun phrases from several texts are presented for the reader to peruse. There is every indication that satisfactory back-of-the-book indexes could be produced automatically, with post-editing to delete superfluous items.

Part II reports on the relationship between English word government and the problem of multiple meaning in material language processing. A set of English words is discussed, each of which has the ability to distinguish among semantic meanings by the use of certain syntactic units such as prepositions. Because prepositions play an important role in making semantic distinctions, a section on prepositional semantics is included.

Unclassified
Security Classification

Unclassified

Security Classification

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
indexing, automatic						
extracting, automatic						
abstracting, automatic						
word government						
natural language processing						
semantic clues						
syntax						
syntactic analysis						
prepositional semantics						

Unclassified

Security Classification