III. DICTIONARY AND HIERARCHY CONSTRUCTION

Claudine Harris

1. The Document Collection

The SMART system is designed to test and evaluate various approaches to document retrieval. The document collection chosen must then possess the following characteristics:

(a) the collection should be large enough to give valid statistical results;

(b) the collection should be short enough to be handled with ease;

(c) the language of the documents should be amenable to inclusion in a dictionary of practicable size;

(d) the semantic content of the documents should lend itself to classification.

With these factors in mind, the Abstracts of Current Computer Literature published in the IRE Transactions on Electronic Computers may constitute a suitable collection. Accordingly, the abstracts published in March, June and September 1959 are the subject of the current experiments. They make up a group of closely allied documents within a restricted field.

The SMART system is capable, at this time, of processing texts up to 1,000 words in length. The given abstracts fall within this range.

2. Nature of the Language Data

Each document consists of a paragraph in the natural language preceded by a title and citation information. At the present time, the bibliographic citation data are not included in the text analysis. Mathematical expressions and proper names of persons, corporate entities or machines occur frequently in the abstracts.

3. The Dictionaries

To allow for flexibility in the retrieval system, dictionaries and tables are provided, each containing linguistic data appropriate to certain phases of the lookup and analysis. These are stem and suffix lists, with their associated semantic and syntactic values; a hierarchical structure of concepts; a list of syntactic criterion trees (phrases); a list of significant word pairs; a simulated vacuous dictionary.

A. Stem and Suffix Dictionaries

The size of the alphabetic dictionary is kept small by entering words in stem form whenever possible. Each entry is supplied with a string of semantic codes which identify the concepts with which it is associated. A second string of codes gives the homographs of the entry for syntactic analysis.

A word is looked up in the stem dictionary by a left-to-right, letter-by-letter comparison. Any suffix present on a text word is then added by scanning the suffix list as many times as is necessary. Suffixes carry no semantic information, but their syntactic codes are combined with the stem syntactic codes. Examples of stem and suffix entries are given in Tables I and II.

Stems	Concept Numbers	Syntactic Codes	Corresponding Partial Homographs
RECKON	0013	043048	OT10,OT60
RECOGN	0332	043048	OT10,OT60
RECORD	01560401	070043	NØUO,OT10
RECORD-REPRODUCE	0401	070	NØUO
RECOVER	00630026	043042	OT10,OI30
RECTI	0306	043	OT10
REDO	0063	043	OT10
REDUCE	01840173	043042	OT10,OI30
REDUCT	01840173	000	0000
REEL	0191	070	NØUO
REFER	0026	042043	OI30,OT10
REFERENCE	0114	043	OT10
REFINE	0063	043042	OT10,OI30
REFLECT	0073	043048	OT10,OT60
REGENERATE	0084	043040070	OT10,OI10,NØUO
REGISTER	0111	070043040	NØUO,OT10,OI10

Typical Alphabetic Stem Dictionary Entries

TABLE 1

The complete homograph for syntactic analysis is obtained by the combination of a stem partial homograph and a suffix partial homograph. For example, a partial homograph such as OT10 will combine with a partial homograph from the suffix list, such as VOOSO, to form a complete homograph. In this case, the complete code is VT1SO, indicating a single object transitive verb in the third person singular.

Alphabetic Suffix List		Syntactic Suffix Codes				
FICATION	058	058	NØUS			
FICATIONS	059	059	NØUP			
FIED	060	060	VOOCO	POO 0	ADJ	
FIER	061	061	NØUS			
FIERS	062	062	NØUP			
FIES	063	063	VOOSO			
FOLD	064	064	ADJ	NØVC		
FUL	065	065	ADJ	NØVC		
FULLY	066	066	AV1			
FY	067	067	VOOPO	IOO 0		
FYING	068	068	ROO 0	GOOSO	NØVS	ADJ

Typical Suffix Dictionary Entries

TABLE 2

When a text word consists of a stem only, without the addition of a suffix, the codes derived from the stem are combined with special codes assigned to "no suffix." These special codes are OOOSO, VOOPO and IOO O, which are capable of combining with noun and verb partial homographs. All other parts of speech carry complete homographs in the stem dictionary, as do irregular verbs and irregular plural nouns.

The present suffix dictionary contains 194 entries. Noun suffixes are entered in plural as well as singular forms, and adjectival suffixes in the adverbial form. Verb suffixes include the common endings "ed," "ing" and "s," as well as true verb suffixes such as "fy" with their inflected forms. This is done for two reasons: to save having to do a double scan of the suffix list and to permit use of the same suffixes in constructing a vacuous dictionary.[†]

[†] Multiple suffixes, such as "fying" could be found by scanning the list first for "fy" then for "ing"; however, the method of construction of the vacuous dictionary does not provide for double scanning of the suffix list, so that the words "identify" and "identifying" would not be recognized as arising from the same stem.

Since the syntactic codes of verb suffixes are partial homographs. it is necessary to enter the combining part of each code as a suitable homograph of any stem capable of taking on a verb suffix. An example of the combination of partial homographs is given in the note to Table 1. The stem "recti" is coded as a potential verb because it can form "rectify" whereas "reduct" carries no syntactic code at all, since it is not a true English word as it stands and can only accept the suffixes "ion" and "ible," which have the complete homographs "NØUS" and "ADJ."

However, in a limited number of cases, partial syntactic coding may introduce an ambiguity: if the word "capital," for example, is coded as a potential verb to accept the suffix "ize," the plural noun "capitals" will receive the extraneous coding of a verb in the third person singular. This difficulty is prevented by entering the stem "capit" with a partial verb code. The suffix "als" properly carries with it only the plural noun code, and "capitalize" is found by a double scan of the suffix list.

Included in the stem dictionary are the so-called common words of the English language which carry no significant semantic values, that is, which do not convey information of value in determining the subject matter of a document. These words are assigned so-called "common" concept numbers to be ignored in the semantic analysis, but their syntactic codes are available for any desired syntactic analysis.

B. The Semantic Concept Hierarchy

Each entry in the stem dictionary is assigned one or more concept numbers. An attempt is made to predict all possible uses of a word and each distinct meaning is coded. Any one concept can contain an unlimited number of stems which may be any part of speech, partial stems, or the mnemonics obtained as output of the phrase-finding routines. Table 3 shows examples of the variety of terms attached to a given concept number.

Concept Number	Examples of Included Terms	Concept Number	Examples of Included Terms
0069	error fail fault mistake	0275	balance compensate
0070	deviate inaccur inexact loss	0280	error-correcting *CORECT (0069,0275) *CORECT (0070,0306) *CORECT (0069,0306)
0215	detect find note	0299	*ERRFND (0069,0215)
		0306	correct fix recti repair

* – mnemonic phrase name with its component concepts

Semantic Concept Classification

TABLE 3

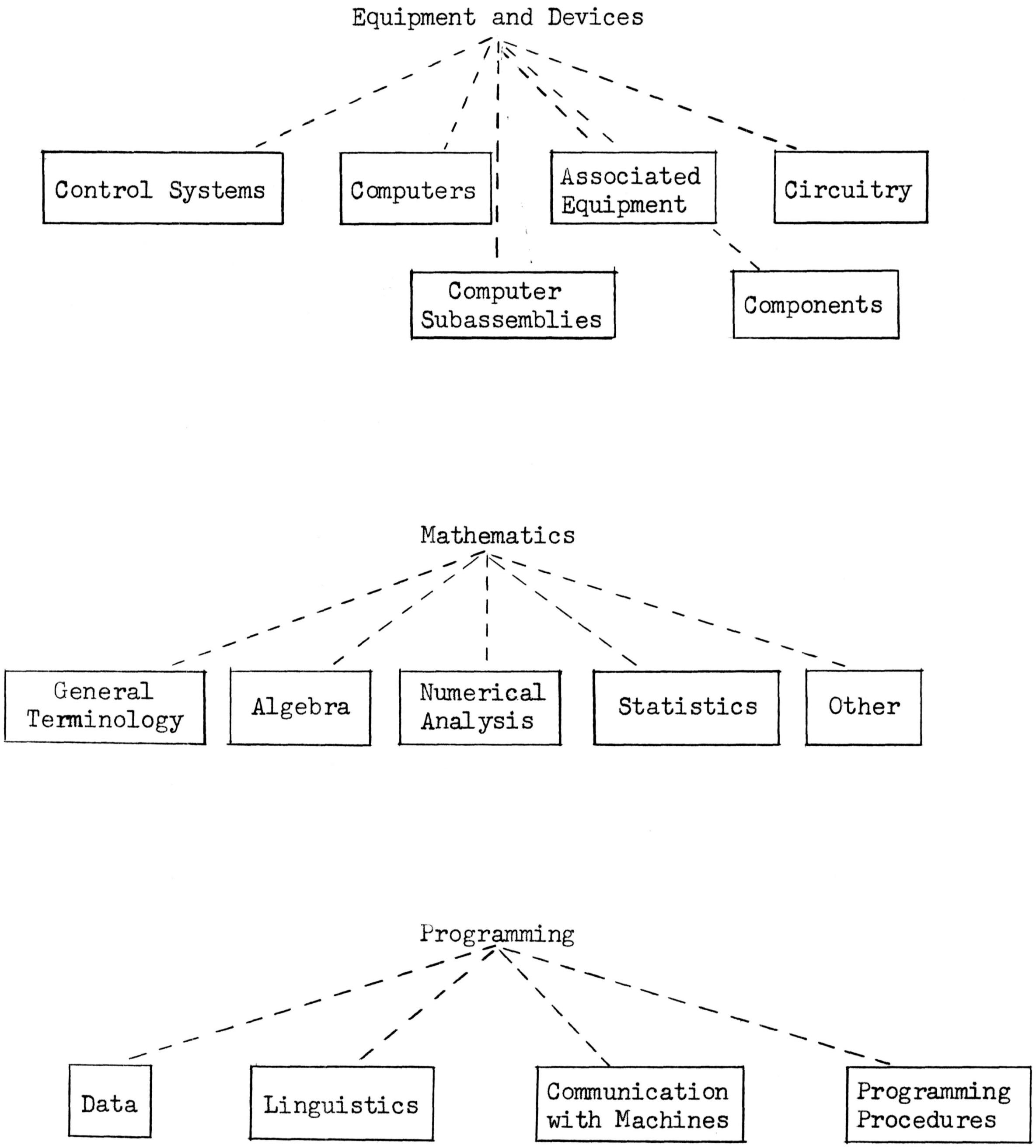
It was first expected to limit the assignment of significant concept numbers to those parts of the natural vocabulary which were immediately apparent as carriers of information. However, a careful reading of the document collection has shown that many obviously technical terms occur also in a general sense, and it becomes necessary therefore to code the significant technical meaning as well as the more general "common" meanings.

The current solution is to consider the natural language as a whole to make up distinct portions of a hierarchy of concepts. It is thus possible to isolate well-defined areas of mathematics, computer technology, or applications. The over-all classification plan is shown in Fig. 1. The inclusion of general meanings in part of the hierarchy introduces difficulties in the statistical analysis of texts, and it is possible that portions of the current dictionary may need to be recoded with "common" meanings and retained only for syntactic analysis.

In the absence of statistical evidence, the technical parts of the concept hierarchy have been structured on an intuitive basis. It is intended that parent nodes contain the more general terms, and filial nodes the more specific terms. The decision to include a term within a certain concept number, or make it a son or a brother is based on the requirement to keep ideas separate. Not enough actual retrieval runs have been made as yet to evaluate specific hierarchical problems. An excerpt of the hierarchy is shown in Fig. 2, corresponding to the concept numbers of Table 3. In order to deep the figure simple, only one term is shown at each node.

C. Phrase Dictionaries – Criterion Trees and Word Pairs

Using the syntactic codes obtained in the look-up procedure, it is possible to perform a syntactic analysis of each sentence and to describe the sentence in syntactic dependency tree form. Typical tree structures are included in a criterion tree dictionary, each node of the tree being assigned the semantic concept numbers of the corresponding words of acceptable natural phrases.



Simplified Outline of the Semantic
Concept Classification

Figure 1

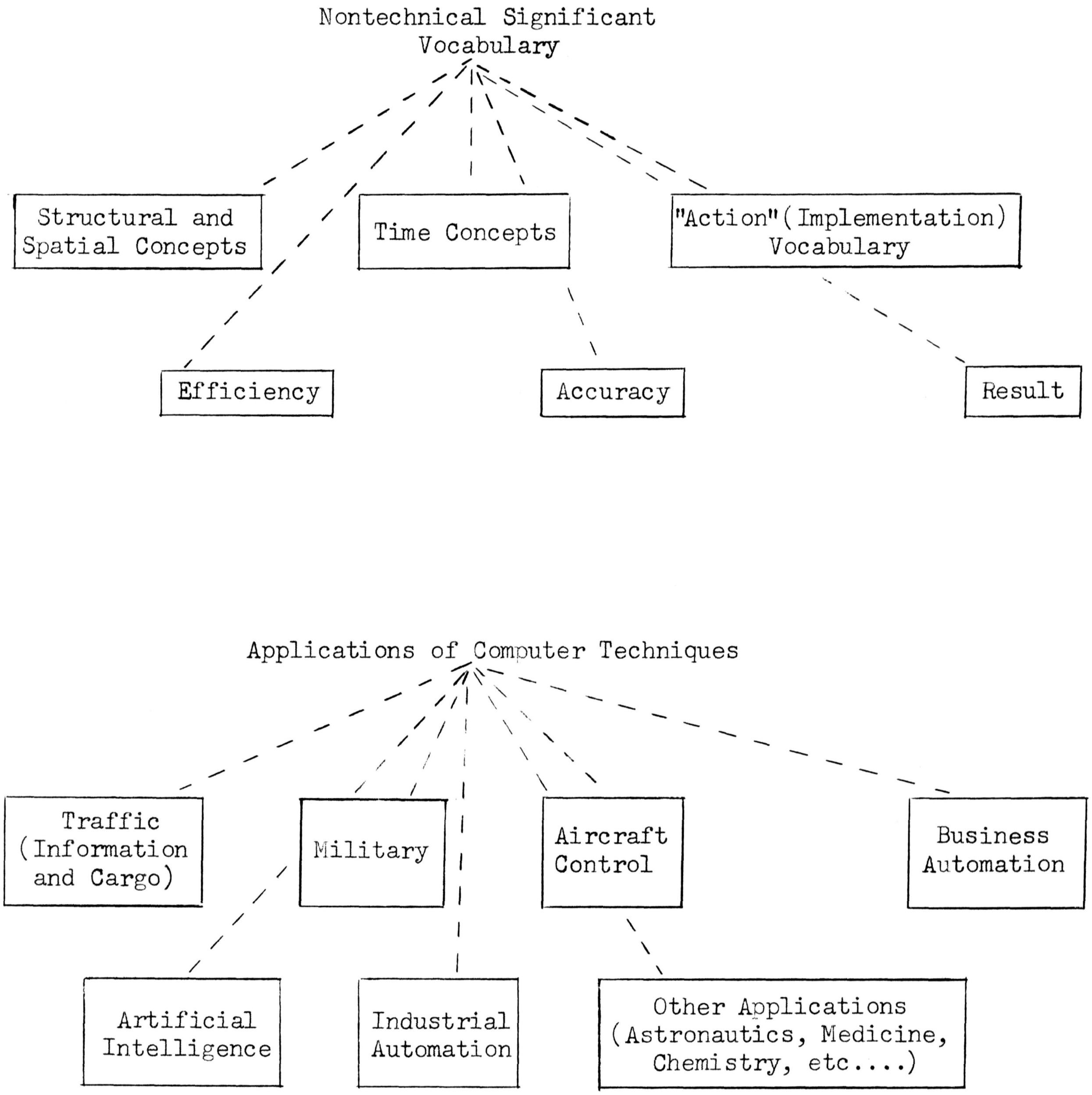
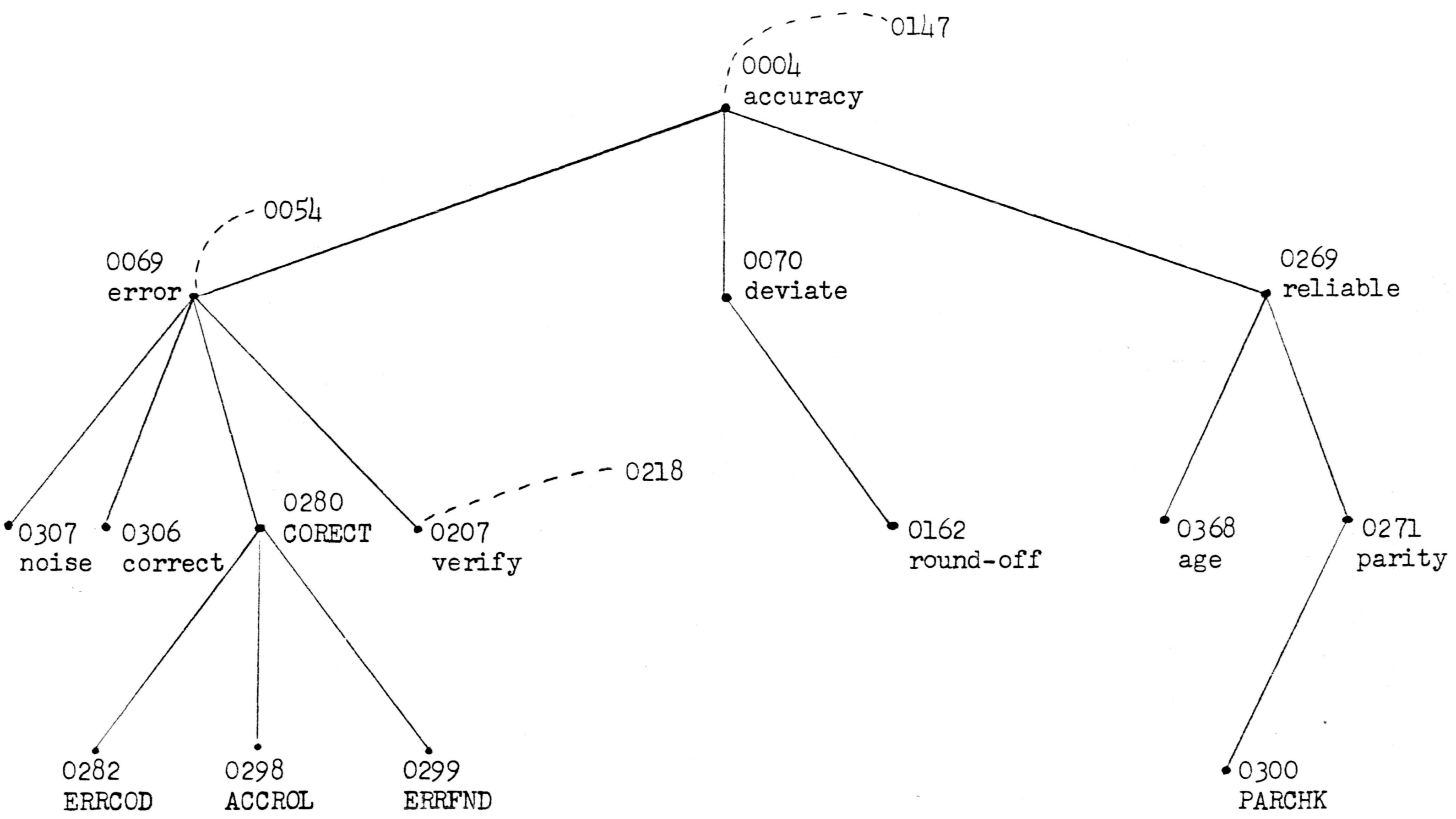

Nontechnical Significant Vocabulary
Structural and Spatial Concepts
Time Concepts
"Action"(Implementation) Vocabulary
Efficiency
Accuracy
Result
Applications of Computer Techniques
Traffic (Information and Cargo)
Military
Aircraft Control
Business Automation
Artificial Intelligence
Industrial Automation
Other Applications (Astronautics, Medicine, Chemistry, etc....)

Figure 1 (continued)



<u>Note</u>: ---- cross references

The following phrase names appear in this excerpt: CORECT, error correction; ERRCOD, error-correcting code; ACCROL, accuracy control; ERRFND, error detection; PARCHK, parity check.

Excerpt from the Concept Hierarchy

Figure 2

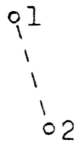
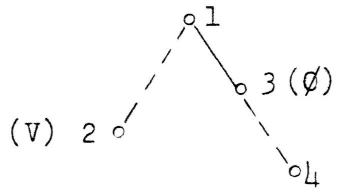
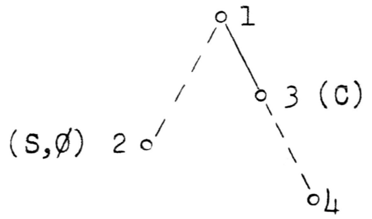
Each tree is given a serial number, a mnemonic phrase name (also called a subject heading), and a semantic concept number. As in the stem dictionary, a single concept number may correspond to several ways of expressing the same idea. There will then be several criterion trees associated with a given concept number, and that number is included in the semantic concept hierarchy.

The list of tree mnemonics is essentially a list of phrase names, or significant word pairs, and it is used in this way by the PHROCC routine. The list of phrase names supplies for each one, in addition to its own concept number, the concept numbers of possible component terms, but carries no syntactic information. Phrase names can therefore be obtained in two ways: as the output of the syntactic criterion tree matching procedure, and as co-occurrence of terms within a sentence irrespective of the syntax. Examples of trees and word pairs are given in Tables 4 and 5.

Phrase Name	Concept Classification	Typical Associated Phrases
CODWRD	0128,0281,0208	code word
CORECT	0280,0275,0069	errors are compensated for
CORECT	0280,0070,0306	correcting inaccuracies is possible
CORECT	0280.0069,0306	error correction is used
DAPROC	0200,0147,0191	tape handling procedures
DAPROC	0200,0147,0053	data processing
DAPROC	0200,0147,0077	the processing of files

Phrase Dictionary Excerpt

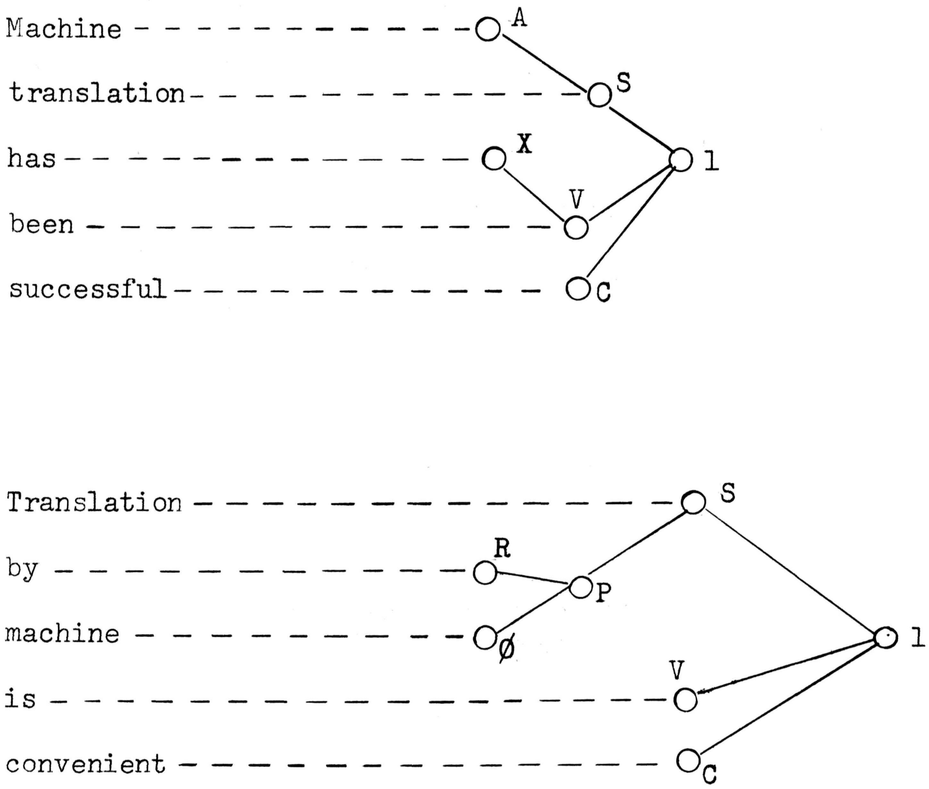
TABLE 4

Subject Heading or Phrase Name	Node Number	Tree Structure	Prescribed Semantic and Syntactic Codes	Node Representation	Serial Number
MCHTRA					140
	1	-0X	(0098)		140
	2	1I	(0119)		140
MCHTRA					150
	1	-0X			150
	2	1I	(0098)(V)		150
	3	1D	(∅)		150
	4	3I	(0119)		150
MCHTRA					160
	1	-0X			160
	2	1I	(0119)(S,∅)		160
	3	1D	(C)		160
	4	3I	(0098)		160

Criterion Tree Dictionary Entries

TABLE 5

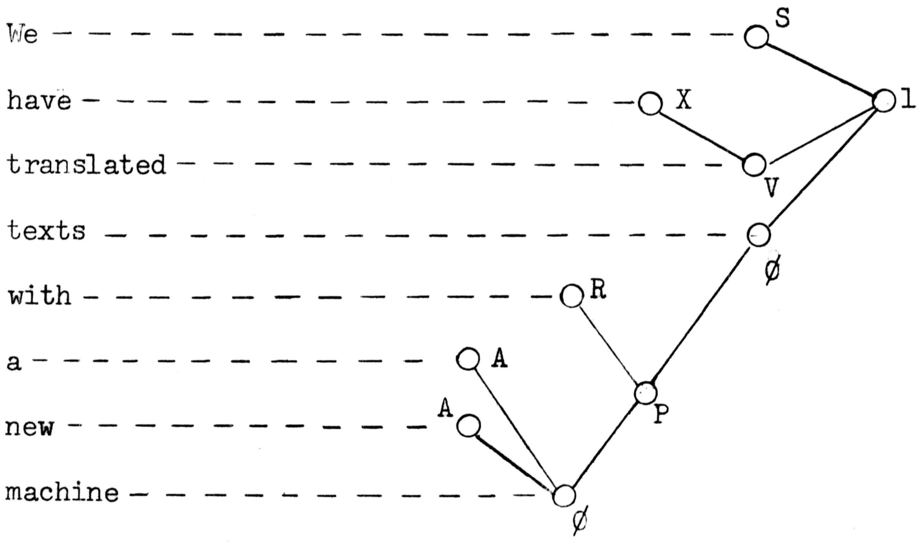
In Table 5, three typical trees are given with a diagram of their node structures. In the first example, the nodes are not restricted as to syntax. This tree will match such sentences as "Machine translation has been successful." and "Translation by machine is convenient." (See Fig. 3.)



Syntactic Analyses Corresponding
to Criterion Tree 140

Figure 3

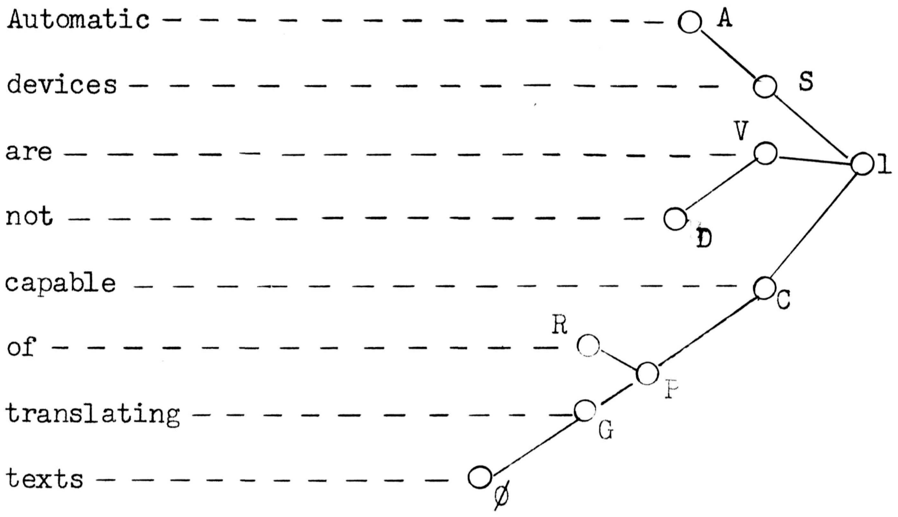
In the second example, node 2 is a verb meaning "translation," node 3 is an object, and node 4 carries the meaning of "machine." A corresponding sentence would be "We have translated texts with a new machine." (See Fig. 4.)



Syntactic Analysis Corresponding
to Criterion Tree 150

Figure 4

The third example specifies that node 2 is either a subject or an object with the meaning of "machine," node 3 is a complement, and node 4 carries the meaning of "translation." A matching sentence would be "Automatic devices are not capable of translating texts." (See Fig. 5.)



Syntactic Analysis Corresponding
to Criterion Tree 160

Figure 5

D. The Vacuous Dictionary

A simulated dictionary can be constructed directly from the input text during the look-up operation. This dictionary assigns a different concept number to each new word that occurs in the text, a new word being defined as any new stem obtained by removing suffixes from a text word with a right-to-left scan. Later analysis of documents is then actually based on the original text words rather than the combined semantic numbers of the concept hierarchy. The suffix list used in constructing the simulated vacuous dictionary is identical with that used for the real stem dictionary.