V. AUTOMATIC PHRASE MATCHING

G. Salton

1. Introduction

In 1957, Luhn suggested a fully automatic procedure for the processing of written texts, based on the frequency of occurrence of words within the texts.[1] Specifically, use of high-frequency words was advocated for purposes of content identification, and document retrieval was to be effected by manipulation of the corresponding word frequency lists. The suggested procedure, while admittedly imperfect, is still used as a basis for many automatic text processing programs.

More recently, the original statistical methods have been modified in various ways: by using word stems rather than the original word forms to identify document content; by introducing synonym dictionaries to lessen the effects of vocabulary variations; and, most importantly, by identifying relations between certain words to be used as content identifiers in conjunction with the surrounding words.

As a result, many of the word matching systems are now being replaced by phrase processing systems, in which the basic units being manipulated are sets of normalized words together with specified relations between them.

In the present report, a variety of methods are described for the automatic generation and manipulation of phrases. The phrase matching procedures used in the SMART document retrieval system to match semantically similar but syntactically quite distinct structures are described in detail, as a specific example of present capabilities.

2. The Content Analysis Problem

In information processing, the structure of written data is of particular importance, because a large part of the information of interest is represented by combinations of words in the natural language. If it is desired to use written data directly as part of an information system, it becomes necessary to define transformations designed to reduce the original input in the natural language into some predetermined standard form. In particular, it would be useful if a text were reducible automatically into a set of controlled terms complemented by a set of well-defined relational indicators.

Before determining the extent to which the known structure of the natural language can help in this endeavor, it may be well to list some of the known difficulties which stand in the way of an automatic content analysis:

(a) the synonym and homograph problem for individual text words (many words can be used to represent the same concept; some words can represent many different concepts);

(b) the problem of semantically equivalent, but syntactically distinct, constructions (a large number of different constructions can be used in the natural language to express the same thought; e.g. "the children broke the window," "the children used rocks to break the window," "rocks were thrown by the children, and as a result the window was broken," etc.);[†]

(c) the problem of indirect reference, including the use of pronouns and other referents to describe information not specifically stated but presumed known from the context (the dependent structures often straddle sentence boundaries in such cases, as in the example:

"Someone opened the door. It was our father.");

(d) the problem of existing relations which are unstated, but may nevertheless be deduced from relationships actually available (in the Syntol system, for example, "associative" relations between a first element and an action, and between the same action and a second element, automatically generate a "consecutive" relation between elements one and two);[‡][4,5,6]

(e) the grouping problem which arises because constructions may refer to a variable number of unspecified items, or to a set of items defined elsewhere (e.g. "all words starting with 'x' are of foreign origin").

It is clear that any one of these difficulties would by itself be sufficient to prevent, in almost all cases, an analysis of written texts into simple components. The presence of homographs and synonyms effectively guarantees that the words used in a text will have to be properly standardized before being used, and the multiplicity of semantically equivalent

[†] A large number of examples of this type are included in Refs. 2 and 3.

[‡] A solution to this problem is sometimes sought in the construction of automatic deductive systems.[7,8]

structures indicates that the isolation of word groupings together with normalized relations between them is likely to be an operation of great difficulty.

The normalization of the vocabulary may be attempted by using a variety of synonym dictionaries and thesauri. Word groupings and relations between words, on the other hand, must be determined in part by utilizing the known structure of the language. This problem is examined in more detail in the next few paragraphs.

3. Language Structure

It is well-known that at least some of the structure of sentences in the natural language is based on syntax, and that this structure is revealed by syntactic analysis. A variety of programs exist to perform automatic syntactic analyses, and these programs are generally based on a form of grammar, known as a context-free phrase structure (type 2) grammar.[9] Such grammars are characterized by the fact that the bracketting used to represent the sentence structure includes both juxtaposed bracket sets as well as nests of brackets, but that interleaving between different bracket pairs is not possible. For example, a structure such as

$$\Big\{\Big\{(A \cdot B) \cdot C\Big\} \cdot \Big[(D \cdot E) \cdot F \cdot \Big\{G \cdot (H \cdot I)\Big\}\Big]\Big]\Big\},$$

where the letters may stand for text words and the bracketting denotes phrase structure, could have been produced by a type 2 grammar. On the other hand, the structure

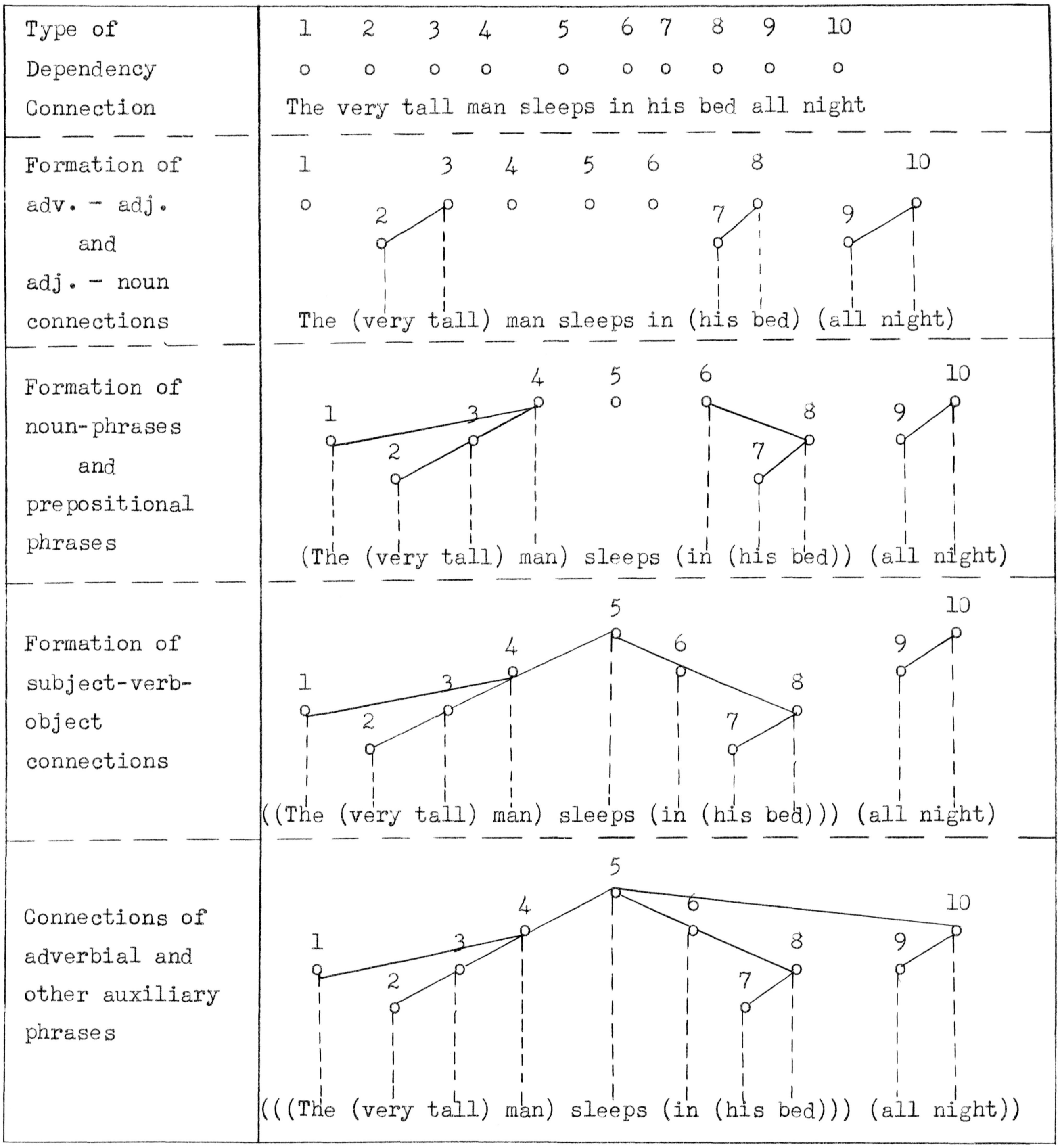
$$A \cdot \{(B \cdot C) \cdot D\}$$

is not producible by such a grammar, because of the interleaving between different bracket pairs.

The syntactic structure of a sample sentence is shown as an example in Fig. 1 in various stages of the analysis. In order to exhibit explicitly the syntactic dependency relations between the words (the relations between governor and dependent words), a dependency model is used, together with the corresponding dependency tree.[†] In Fig. 1, a word A appearing below another word B to which it is attached by a branch is syntactically dependent on it; furthermore the bracketting structure is given directly by the subtree arrangement in the dependency tree (or in the corresponding phrase structure tree).

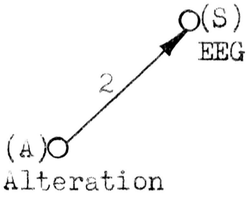
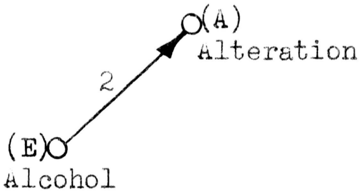
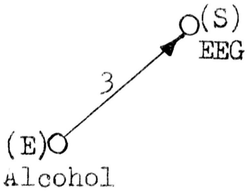
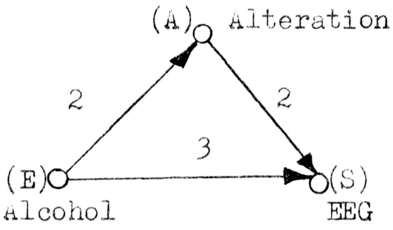
Phrase structure analyses are of particular interest in the present context not only because a variety of machine programs exist which can perform such analyses automatically,[11] but also because phrase structure, as the name indicates, accounts for the most important word groupings, including noun phrases, prepositional phrases, adverbial phrases, and in most cases for the basic subject-verb-object grouping. These groups are also those which make up the basic components to be included in a useful information graph, as seen for example in Fig. 2, illustrating the (manual) construction of a Syntol graph.

[†] Dependency grammars have been shown to be (weakly) equivalent to phrase structure grammars; the two models can be used interchangeably for present purposes.[10]


Type of Dependency Connection
1 2 3 4 5 6 7 8 9 10
The very tall man sleeps in his bed all night
Formation of adv. - adj. and adj. - noun connections
The (very tall) man sleeps in (his bed) (all night)
Formation of noun-phrases and prepositional phrases
(The (very tall) man) sleeps (in (his bed)) (all night)
Formation of subject-verb-object connections
((The (very tall) man) sleeps (in (his bed))) (all night)
Connections of adverbial and other auxiliary phrases
(((The (very tall) man) sleeps (in (his bed))) (all night))

Formation of Dependency Connections for a Sample Sentence

Figure 1

Tree or Graph Form	English Correspondent
	Alteration of the EEG (partial information)
	Alcohol alters something (partial information)
	Alcohol affects the EEG (partial information)
 (A) Alteration 2 2 3 (E) Alcohol (S) EEG	Alteration of the EEG due to alcohol (complete information item)

2	Associative relation
3	Consecutive relation
A	Action
E	Entity
S	State

Construction of a Typical Information Graph

Figure 2

Some phrases or word groups whose component words do not occur in adjacent word positions within a sentence are difficult to generate by an unmodified phrase structure grammar. This is the case notably of phrases with the so-called discontinuous constituents (e.g. "call up" in "call him up"). In order to accommodate discontinuous constituents, the normal type 2 production rules must be extended, thus tending to produce a relatively complicated grammar.[11,12]

Some important linguistic phenomena do not fit into a phrase structure model, even if extended to handle special cases such as discontinuous constituents. There is no way in a phrase structure model to relate, for example, two semantically identical sentences of which one is in the passive and the other in the active voice. It is often suggested, therefore, that in order to produce correct word groupings for a variety of transformed structures, a transformational grammar be added to the phrase structure model. Such a move could be expected to produce not only a grammar more nearly representative of natural language structure as it exists, but would also result in a simpler, more economical, phrase structure component.

A possible procedure advocated for an automatic sentence analysis, and for the generation of basic word groups, or kernels, consists in the alternate application of phrase structure rules and inverse transformations. Specifically, phrase structure rules are applied first to produce a standard phrase structure analysis of an input string; the analyzed string is then subjected to all applicable inverse transformations. The transformed strings are then analyzed once more, and so on, until no further change is produced in the output.[13,14]

This procedure may be expected to produce a much larger number of correct word groupings than can be obtained from a phrase structure analysis alone. On the other hand, the apparatus required to use a transformational grammar as part of an automatic system may be expected to be much more complex than the simple pushdown store analysis, or list-tracing procedures, needed to use a simple phrase structure grammar. Whether the combined phrase structure and transformational procedure turns out to be effective in the generation of word groups needed for content description remains to be seen. In any case, experimental kernelizing programs are presently under study by several research groups.[3,15,16]

4. The Processing of Unrestricted Text

A number of text processing systems have been designed to process completely unrestricted natural language input. Among these are at least two systems designed to answer questions rather than to furnish references, the Protosynthex and the FLEX systems.[17,18]

In order to be able better to assess the problems raised by systems such as the two previously mentioned, it is convenient to consider some of the text processing methods included in a fully automatic document, rather than fact, retrieval system. The SMART system[19,20] takes both documents and search requests in unrestricted English, performs a complete content analysis automatically, and retrieves those documents which most nearly match the given search request. A large variety of procedures are available for the generation of the content identifiers attached to both search requests and stored documents, and documents may therefore be retrieved in accordance with many different criteria.

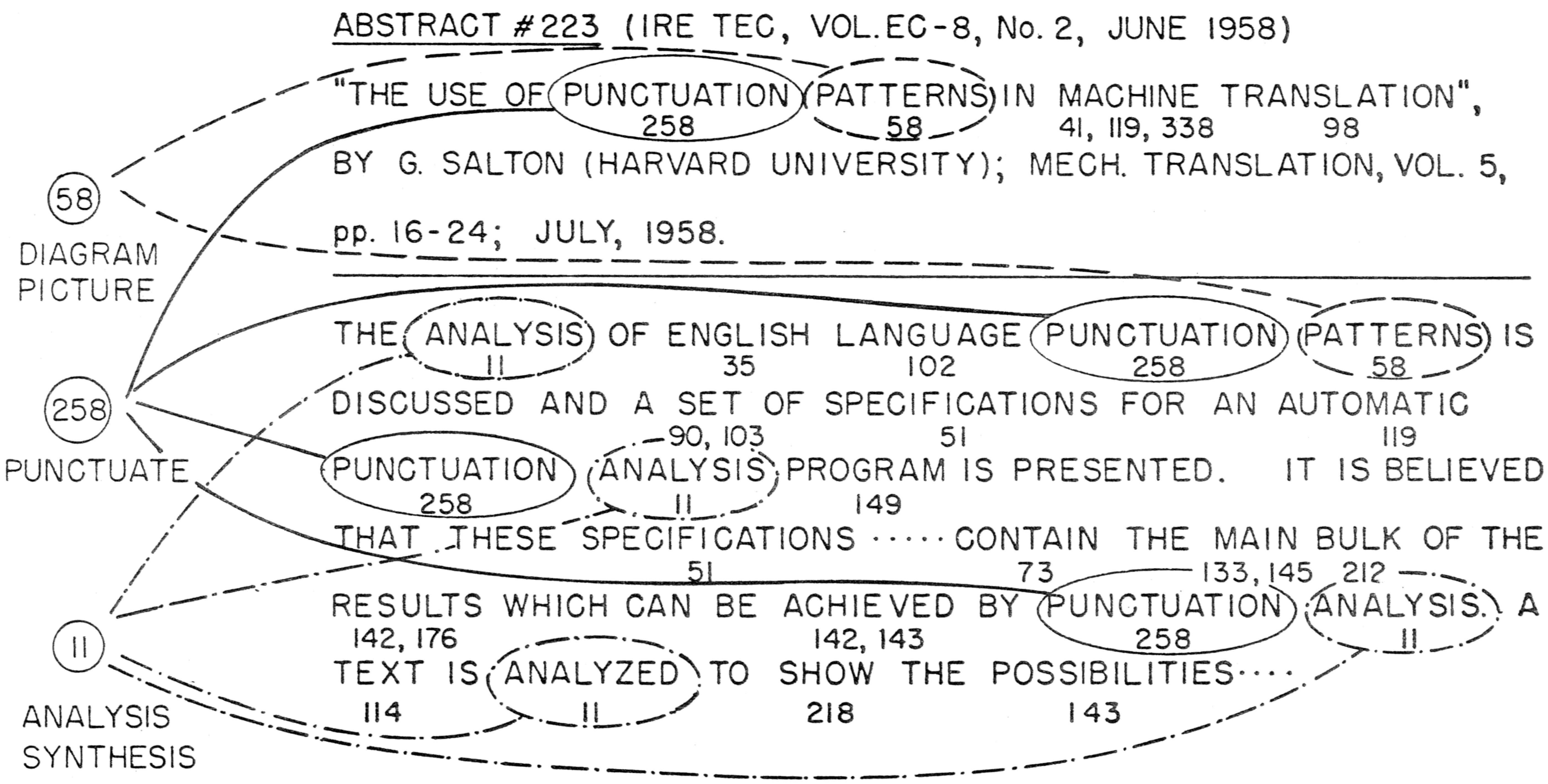
The system can be controlled by the user in that a search request is first processed in a standard mode; the user is then asked to analyze the output obtained, and depending on his further requirements, the original search request can be reprocessed using a new processing method. The new output is then again examined and the process can be iterated until such time as the right kind and amount of information are retrieved. The various processing modes correspond to different automatic methods of analyzing information, and the iterative procedure represents an attempt to approximate, with natural language input, the type of analysis (in terms of controlled concepts together with controlled relations between them) previously illustrated by the manual process of Fig. 2. Before exhibiting the differences between the theoretically desirable and the actually achievable reduction, some of the basis SMART operations are outlined briefly.

The first operation consists generally in a stem-suffix cut-off operation, which replaces each text word occurring in a document or in a search request by the corresponding word stem. High-frequency function words, such as conjunctions, prepositions, and the like may then be temporarily discarded, and a document (or a search request) can be identified by the set of remaining word stems, together with a frequency indicator for each stem. At this point, the word stems used to represent item properties are not as yet normalized.

In order to reduce synonymous word stems to a single "concept," and to provide a variety of different concept identifications for the many stem homographs which may arise in the natural language, it is necessary to perform a thesaurus look-up operation. This process effectively replaces each word stem by one or more so-called concept numbers. The replacement of word stems by concept numbers is illustrated in Fig. 3 for a typical document abstract included in SMART. Use of the thesaurus insures that a given document is identified by a set of controlled terms.

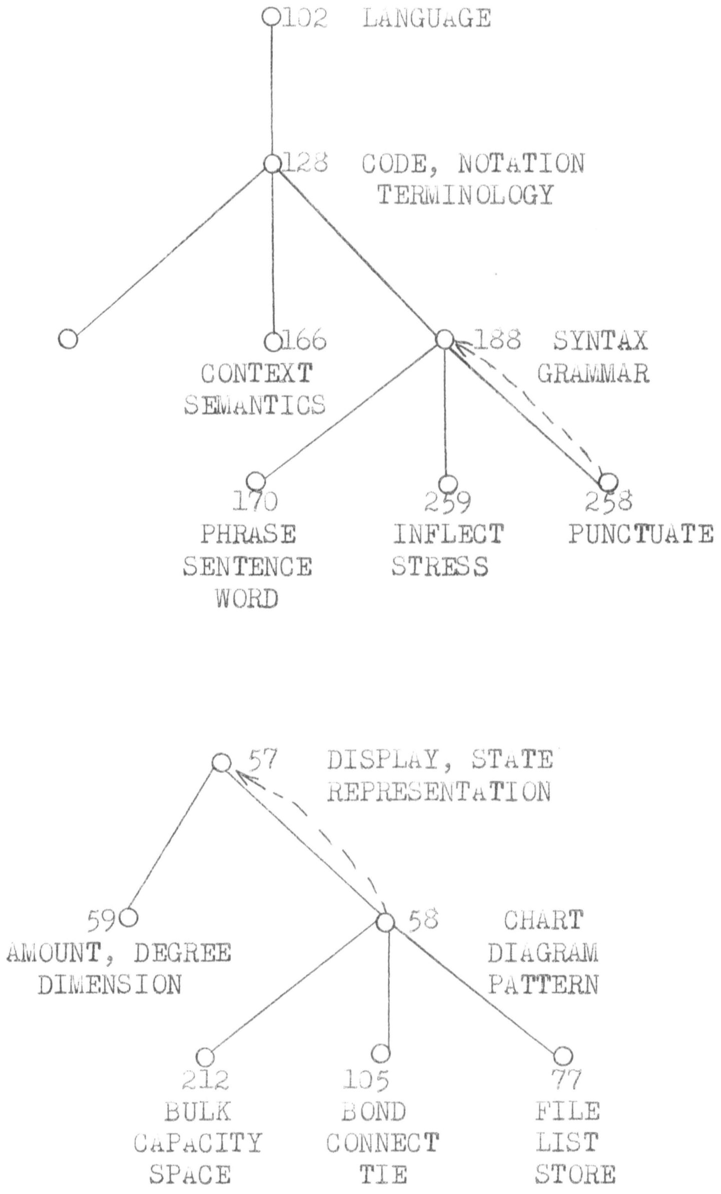
Generic relations between properties may be provided by consulting a hierarchical arrangement of concept numbers as shown in Fig. 4. Specifically, given any concept obtained from the thesaurus, it is now possible to obtain related concepts by using the tree structures. More general concepts may be located by going "up" in the tree, more specific ones by going "down," and various related concepts may be picked up by locating the "brothers" (nodes in the same filial set) of a given concept. Figure 4 illustrates, for example, the expansion of concept 258 (punctuation) into concept 188 (syntax), and of 58 (pattern) into 59 (representation).

Relations between concepts may be added by grouping the concepts instead of using them one at a time. The identification of so-called statistical phrases is illustrated in Fig. 5 for the document abstract previously shown in Fig. 3. Statistical phrases are groups of concepts


ABSTRACT #223 (IRE TEC, VOL. EC-8, No. 2, JUNE 1958)
"THE USE OF PUNCTUATION PATTERNS IN MACHINE TRANSLATION",
258 58 41, 119, 338 98
BY G. SALTON (HARVARD UNIVERSITY); MECH. TRANSLATION, VOL. 5,
pp. 16-24; JULY, 1958.
THE ANALYSIS OF ENGLISH LANGUAGE PUNCTUATION PATTERNS IS
11 35 102 258 58
DISCUSSED AND A SET OF SPECIFICATIONS FOR AN AUTOMATIC
90, 103 51 119
PUNCTUATION ANALYSIS PROGRAM IS PRESENTED. IT IS BELIEVED
258 11 149
THAT THESE SPECIFICATIONS ····· CONTAIN THE MAIN BULK OF THE
51 73 133, 145 212
RESULTS WHICH CAN BE ACHIEVED BY PUNCTUATION ANALYSIS. A
142, 176 142, 143 258 11
TEXT IS ANALYZED TO SHOW THE POSSIBILITIES····
114 11 218 143
58
DIAGRAM
PICTURE
258
PUNCTUATE
11
ANALYSIS
SYNTHESIS

Assignment of Concept Numbers

Figure 3



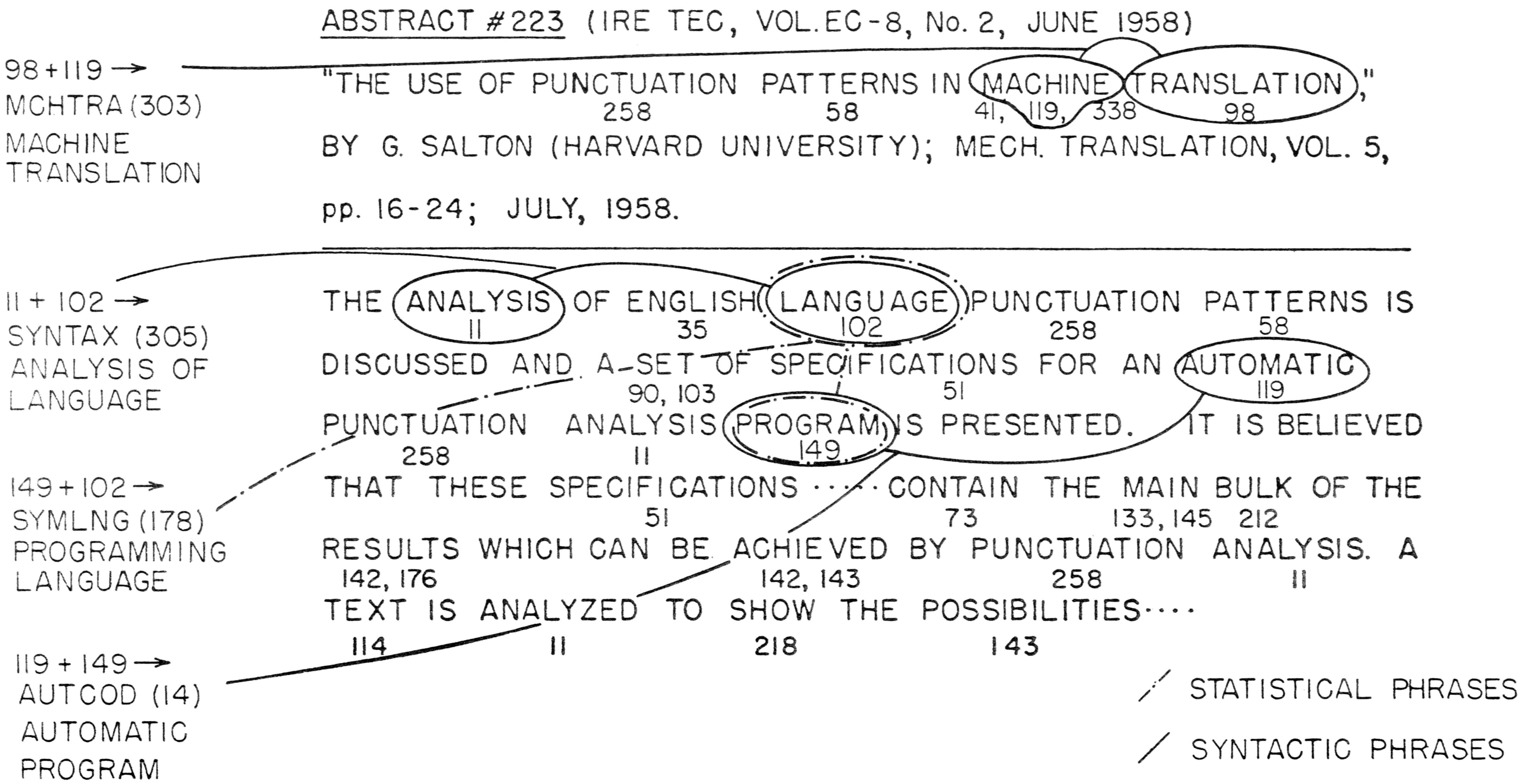
Hierarchy Excerpt Showing Expansion

Figure 4

which co-occur within the sentences of the documents with a frequency exceeding some pre-established threshold. If such a group of concepts is detected, the individual concept numbers may be replaced by a group concept number attached to the statistical phrase; such a phrase concept may then be given a higher weight than the individual word concepts when it is used as part of a document identification.

In the abstract of Fig. 5, co-occurrence in the same sentence of concepts 11 (analysis) and 102 (language) is recognized as a phrase with concept number 305 (language analysis); similarly for concepts 102 (language) and 149 (program), which are transformed into 178 (programming language). The exact type of relation which obtains between a concept pair included in a given statistical phrase cannot in general be determined, since the formation of such phrases depends strictly on concept co-occurrence characteristics. Thus, the use of statistical phrases does not completely fill the requirements of the graph model of Fig. 2, where relations are identified completely.

An attempt to identify at least some relations between concepts may be made by using the syntax of the language. Specifically, statistical phrases may be replaced by syntactic phrases in which the included concepts exhibit some specified syntactic relationship. In order to be able to include syntactic relationships as part of the content identification process, it is necessary to perform an automatic syntactic analysis. Such a step is included in the SMART system, and procedures are provided for eliminating statistical phrases which do not qualify as proper syntactic phrases. In the abstract of Fig. 5, for example, the statistical phrase


ABSTRACT #223 (IRE TEC, VOL.EC-8, No. 2, JUNE 1958)
"THE USE OF PUNCTUATION PATTERNS IN MACHINE TRANSLATION,"
258 58 41, 119, 338 98
BY G. SALTON (HARVARD UNIVERSITY); MECH. TRANSLATION, VOL. 5,
pp. 16-24; JULY, 1958.
THE ANALYSIS OF ENGLISH LANGUAGE PUNCTUATION PATTERNS IS
11 35 102 258 58
DISCUSSED AND A SET OF SPECIFICATIONS FOR AN AUTOMATIC
90, 103 51 119
PUNCTUATION ANALYSIS PROGRAM IS PRESENTED. IT IS BELIEVED
258 11 149
THAT THESE SPECIFICATIONS ···· CONTAIN THE MAIN BULK OF THE
51 73 133, 145 212
RESULTS WHICH CAN BE ACHIEVED BY PUNCTUATION ANALYSIS. A
142, 176 142, 143 258 11
TEXT IS ANALYZED TO SHOW THE POSSIBILITIES····
114 11 218 143
98 + 119 →
MCHTRA (303)
MACHINE TRANSLATION
11 + 102 →
SYNTAX (305)
ANALYSIS OF LANGUAGE
149 + 102 →
SYMLNG (178)
PROGRAMMING LANGUAGE
119 + 149 →
AUTCOD (14)
AUTOMATIC PROGRAM
STATISTICAL PHRASES
SYNTACTIC PHRASES

Phrase Matching

Figure 5

corresponding to concept 178 (programming language) is not a syntactic phrase, since an admissible syntactic relation does not exist between the included concepts 102 (language) and 149 (program).

Identification of syntactic relationships between concepts included in a phrase does furnish some relational indications in accordance with the requirements of the model of Fig. 2. However, in order to generate a completely defined graph structure, both concepts and relations must be properly normalized. That is, different syntactic structures must be transformed into the same phrase if there exists semantic equivalence. To what extent this can be done automatically is further described in the next few paragraphs.

The complete content analysis process is summarized in Fig. 6 for the document abstract previously shown in Figs. 3 and 5.

5. Syntactic Phrase Matching

The identification of information items by concept numbers and phrases of various kinds is of use only if the corresponding identifiers are, in fact, properly normalized. This is achieved in the SMART system by replacing words by concept numbers, by performing a syntactic analysis of the sentences occurring in documents and search requests so as to determine syntactic dependency relations between concepts, and finally by looking up the resulting dependency tree structures (see Fig. 1) in a dictionary of <u>criterion trees</u>.

Criterion trees consist of prestored frames including concept numbers, syntactic indicators, and the syntactic dependency relations

Method of Generation	Identifiers	
	Concept Numbers	English Examples (weights)
Thesaurus Look-up	11	Analysis, Synthesis (48)
	58	Diagram, Picture (24)
	59	Amount, Extent (12)
	98	Translation (12)
	102	Language (12)
	119	Automatic, Machine (24)
	149	Program, Routine (12)
	170	Phrase, Word, Sentence (12)
	258	Punctuation (48)
Hierarchy Expansion	57 (58,59)	Display, Represent (36)
	112 (119)	Manipulate, Operate (24)
	188 (170,258)	Grammar, Syntax (60)
Statistical Phrases	178 (149,102)	SYMLNG (program, language)
	305 (11,170)	SYNTAX (analysis, word)
Syntactic Phrases	303 (119,98)	MCHTRA (machine translation)
	305 (11,102)	SYNTAX (analysis of language)
	14 (119,149)	AUTCOD (automatic program)

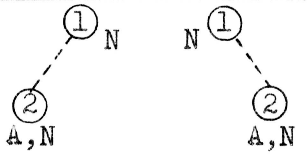
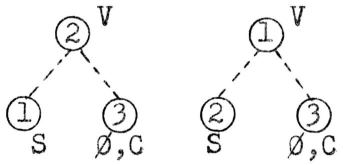
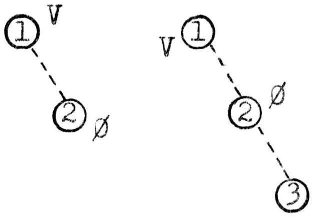
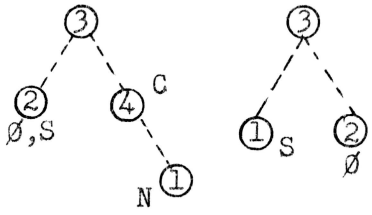
Identification of Abstract #223

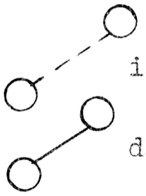
Figure 6

which obtain between the included concepts. There exist four main classes of criterion trees as shown in Fig. 7, corresponding to noun phrases, subject-verb relations, verb-object relations, and subject-object relations. The syntactic structure of criterion trees is conveniently specified by syntactic dependency trees as seen in the center section of Fig. 7. The top-most tree in Fig. 7 corresponds, for example, to an English phrase consisting of an adjective (A), or noun (N), (tree node 2) which is syntactically dependent upon another noun specified by tree node 1.

The operations of the criterion tree dictionary may be explained by considering the example of Fig. 8. The tree, termed SYNTAX, is defined at the top of the figure. Node 1 of the tree must correspond to either concepts 11 or 158, and node 2 to concepts 102, 188, or 170. Furthermore, four different syntactic frames are allowed for the tree, as indicated by the format numbers which follow the $ sign (fourteen different formats are used at present in the criterion tree dictionary). A few typical word stems corresponding to the concepts included in the SYNTAX tree specification are also shown in Fig. 8, as are examples of English phrases and sentences which will match the given tree.

Obviously, the multiplicity of concepts attached to a given node of a criterion tree, and the variety of permissible syntactic formats guarantees that a given criterion tree specification corresponds to hundreds of different English constructions. Furthermore, both documents and search requests use the same criterion tree dictionary, so that a

Tree Type	Dependency Trees	English Examples
Noun Phrases or Prepositional Phrases		Syntactic analysis (2) (1) Analysis of phrases (1) (2)
Subject-verb		This machine translates... (1) (2) the translation appears... (2) (1)
Verb-object Verb-complement		We translate texts (1) (2) by machine (3)
Subject-object Subject- complement		data are available for (2) processing (1) machines perform translation (1) (2)



indirect connections

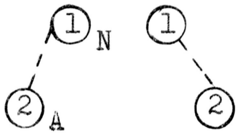
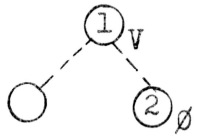
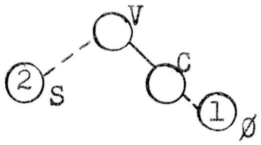
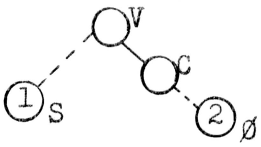
direct connections (not shown)

Basic Criterion Tree Classes

Figure 7

PHRASE SPECIFICATION:

$$\text{SYNTAX } \underbrace{(11,158)}_{\text{CONCEPT NODE 1}}/\underbrace{(102,188,170)}_{\text{CONCEPT NODE 2}}\ \emptyset\ \underbrace{1,3,4,13}_{\text{FORMATS}}$$

NODE 1	NODE 2	FORMATS	SAMPLE PHRASES
11 ANAL SYNTHESIS SYNTHES SYNTHET 158 CLASS CORRESPOND GROUP INDEPEND RELATE	102 INTERLINGU LANGUAGE 170 PHRASE SENTENCE SUBJECT WORD 188 GRAMMAR SYNTAX SYNTACTIC	1 	1 SYNTACTIC ANALYSIS PHRASE RELATIONS ANALYSIS OF SENTENCES
		3 	3 WE CAN ANALYZE THE LANGUAGE ...SYNTHESIZE A SYNTAX
		4 	4 THE GRAMMAR IS NOW AVAILABLE FOR ANALYSIS
		13 	13 THIS ANALYSIS IS APPLICABLE TO RUSSIAN GRAMMAR

Criterion Phrase Specifications

Figure 8

flexible matching process ensues. The comparison of concept numbers and of syntactic indicators is done in the SMART system by table look-up, and the dependency structures of input sentences and criterion trees are compared by an efficient graph-matching process.[21,22] The Kuno-Oettinger multiple path syntactic analyzer is used to perform the automatic syntactic analysis of the input documents.[†][23]

In order to evaluate to what extent the automatic criterion tree procedures can approximate the manual analysis specified by the model of Fig. 2, it is of interest to examine in more detail the variety of different structures which can be matched. The flexibility of the procedure arises from four principal characteristics of the criterion trees:

(a) word stems rather than complete words are used during the thesaurus look-up;

(b) concept numbers rather than words or word stems are attached to the criterion tree nodes;

(c) each criterion tree is assigned several possible syntactic frames, or equivalently, a variety of syntactic relations are normally allowed between concepts; and

(d) the dependency connection between two specified concept numbers is an <u>indirect</u> connection, ignoring occurrences of extraneous concepts, or of function

[†] It is obvious that if the syntactic analysis procedure furnishes an incorrect or a doubtful analysis, the phrase matching process may be affected adversely.

> words which may be part of the syntactic context (thus, the preposition "of" in the phrase "retrieval of information" is ignored when the dependency trees are matched).[†]

As a result of this type of criterion tree specification, it is in general possible to match semantically equivalent phrases or sentences, provided that the same basic order between major sentence parts (subject, verb, object) is present. Differences due to addition or deletion of auxiliary particles and phrases, shifts from noun to verb constructions or vice-versa, use of synonyms and of multiple subjects, verbs, or objects do not in general interfere with the matching procedure. Examples of various syntactic constructions which can be properly recognized by the criterion tree procedure are given in Fig. 9.[‡] Within each group, the sample phrases match the same basic criterion tree.

Since function words, including prepositions, adverbs, and conjunctions, are not normally included in a criterion tree specification, a variety of structures which are not completely synonymous are nevertheless assumed to be identical by the criterion tree routine. Typical examples of the nonrecognition of semantic differences, as well as some examples of the nonrecognition of semantic similarities are shown in Fig. 10. Figure 10 may in fact be considered to be a repertoire of the

[†] This feature is also incorporated in a somewhat different form in a number of other text processing systems.[24]

[‡] Some of the examples included in Figs. 15 and 16 were suggested in studies dealing with the construction of transformational grammars.[25,26]

Transformations	Examples of
Correctly Identified	Matching Structures
1. declarative vs. interrogative (word order between principal sentence parts is maintained)	the man bats the ball does the man bat the ball the boy asks whether the man bats the ball
2. identification of multiple subjects, verbs, objects	the large, grey, empty hall... the large hall... the grey hall...
3. "there is" or "it is" constructions	the car is in the garage there is a car in the garage
4. deletions of certain pronouns	this is the information that you wanted this is the information you wanted
5. permutations within noun and prepositional phrases	pattern analysis... the analysis of patterns...
6. some negative constructions	children do not like teachers no children like teachers
7. identification of synonymous constructions	the grammar of this coding system... the syntax of this notation...
8. identification of stem similarities	analyzer... analysis...
9. verb-noun shifts	he dances; he is a dancer he looks; he gives a look
10. addition of subject or object clauses	the boy works the father demands that the boy should work
11. certain equivalent constructions	how much...; what... which time...; at what time... for more than...; for longer than...

Correctly Identified Phrase and Sentence Transformations

Figure 9

Type of Deficiency	Examples {matching} ≠ not matching ≠
1. active-passive changes not immediately identified (due to change in basic structure)	≠ the man hits the ball ≠ the ball is hit by the man ("man eats dog," "dog eats man" are, however, distinguished)
2. no distinction between depth of dependency connection	{analysis of English patterns analysis of English analysis of patterns}
3. no recognition of negative-positive transformation	{the sun shines the sun does not shine}
4. no recognition of some relative clauses causing word order changes	≠ I saw the man ≠ It is he whom I saw
5. no recognition of dependencies across sentence boundaries	≠ Mr. X is tall. He is our teacher. ≠ Mr. X is our teacher.
6. no recognition of unstated classifications	≠ This poodle is big ≠ This poodle is a big dog ≠ They are 1000 feet apart ≠ The distance between them is 1000 feet.
7. no recognition of distinct verb forms	{The data are retrieved The data were retrieved The data have not been retrieved}

{ nonrecognition of semantic differences

≠ nonrecognition of semantic similarities

Deficiencies in Phrase Matching Process
Figure 10

deficiencies of the SMART phrase matching process. The nonrecognition of the semantic differences illustrated in examples 2, 3, and 7 of Fig. 10 is generally of no consequence for document retrieval, and may also be of trivial importance in question-answering systems based on the given phrase matching process. On the other hand, the nonrecognition of some of the semantic similarities, notably those illustrated by examples 1 and 5, may be expected to be serious, at least for automatic question-answering.

To summarize, the criterion tree matching routine can be used in automatic text processing systems to furnish groupings between specified concepts and to identify a limited number of syntactic dependency relations. While the resulting structures do not completely obey the specifications of the graph model of Fig. 2, the identification obtained is sufficiently detailed to handle satisfactorily the great majority of the problems arising in automatic document retrieval. Additional work remains, however, to be done, including the construction of workable kernelizing routines, before fully automatic, unrestricted question-answering systems become feasible.

REFERENCES

1. Luhn, H. P., "A Statistical Approach to Mechanized Encoding and Searching of Literary Information," IBM Journal of Research and Development, Vol. 1, No. 4 (October 1957).

2. Robinson, J. J., "Preliminary Codes and Rules for the Automatic Parsing of English," Rand Memorandum RM-3339-PR, Rand Corporation (December 1963).

3. Robinson, J. J., "Automatic Parsing and Fact Retrieval: A Comment on Grammar, Paraphrase, and Meaning," Rand Memorandum RM-4005-PR (February 1964).

4. Gardin, J. C., and Levy, F., "Le Syntol - Syntagmatic Organization Language," Proceedings IFIP Congress-62, North Holland Publishing Company (1962).

5. Gardin, J. C., et al., "Final Report on a General System for the Treatment of Documentary Data (Theoretical Applications of Syntol - Part I, Programming of Syntol - Part II)," Association Marc Bloch (October 1963).

6. Coyaud, M., "Analyse Automatique Syntol," presented at the NATO Advanced Study Institute on Automatic Documentation, Venice (July 1963).

7. Black, F., "A Deductive Question-Answering System," Doctoral Thesis, Harvard University (June 1964).

8. Cooper, W. S., "Fact Retrieval and Deductive Question-Answering Information Retrieval Systems," Journal of the ACM, Vol. 11, No. 2 (April 1964).

9. Chomsky, N., Syntactic Structures, Mouton and Co., s'Gravenhage (1957).

10. Hays, D. G., "Dependency Theory: A Formalism and Some Observations," Report RM-4087-PR, The Rand Corporation (July 1964).

11. Bobrow, D. G., "Syntactic Analysis of English by Computer - A Survey," Proceedings AFIPS Fall Joint Computer Conference, Vol. 24, Las Vegas (1963).

12. Yngve, V. H., "Computer Programs for Translation," Scientific American (June 1962).

13. Petrick, S. R., "On Transformational Grammar Recognition Procedures," unpublished manuscript.

14. Herzberger, H. G., "Some Aspects of Kernelization," Western Reserve University (November 1963).

15. Carmody, B. T., and Jones, P. E., Jr., "Automatic Derivation of Constituent Sentences," 1964 Annual Meeting of the Association for Machine Translation and Computational Linguistics, Indiana University (1964).

16. Bobrow, D. G., "Natural Language Input for a Computer Problem Solving System," MIT Report MAC-TR-1 (September 1964).

17. Simmons, R. F., Klein, S., and McConlogue, K., "Indexing and Dependency Logic for Answering English Questions," American Documentation, Vol. 15, No. 3 (July 1964).

18. Thorne, J. P., "Automatic Language Analysis," Final Technical Report to RADC under contract AF 30 (602) - 2185, Indiana University (December 1962).

19. Salton, G., et al., Information Storage and Retrieval, Report No. ISR-7 to the National Science Foundation, Computation Laboratory of Harvard University (June 1964).

20. Salton, G., "A Document Retrieval System for Man-machine Interaction," Proceedings of the ACM 19th National Conference, Philadelphia (1964).

21. Sussenguth, E. H., Jr., "Structure Matching in Information Processing," Doctoral Thesis, Information Storage and Retrieval, Report No. ISR-6 to the National Science Foundation, Computation Laboratory of Harvard University (April 1964).

22. Salton, G., and Sussenguth, E. H., Jr., "Some Flexible Information Retrieval Systems Using Structure Matching Procedures," Proceedings of the AFIPS Spring Joint Computer Conference, Washington (April 1964).

23. Kuno, S., and Oettinger, A. G., "Multiple-Path Syntactic Analyzer," Proceedings of the IFIP Congress-62, North Holland Publishing Company (1962).

24. Klein, S., and Simmons, R. F., "Syntactic Dependency and the Computer Generation of Coherent Discourse," Mechanical Translation, Vol. 7, No. 2 (August 1963).

25. Harris, Z. S., English Transformation List, Transform and Discourse Analysis Project, Paper No. 30, University of Pennsylvania (1964).

26. Hall, B., "Notes on Transformational Grammars," unpublished manuscript.