Automated Language Processing

G. Salton

Technical Report
No. 68-6

February 1968

Department of Computer Science
Cornell University, Ithaca, N. Y.

Automated Language Processing

G. Salton*

Abstract

This study covers recent developments in automatic text processing, including syntactic, semantic, and statistical language analysis methods. The emphasis is on applications in the areas of machine translation, information retrieval, and question answering. Recent on-line text processing methods, using man-machine interaction are covered in some detail, as are certain simple, syntactic analysis methods incorporated into a number of experimental information retrieval systems. An evaluation is made of the syntactic analysis methods, and their importance in a retrieval environment is discussed.

1. Introduction

The present study on automated language processing is a somewhat subjective view of the field designed to reflect the state of the art as of the end of the calendar year 1967. It is subjective because a choice had to be made about what should be covered, and an orientation had to be given to the material actually discussed. Since an excellent recent survey already exists covering the more theoretical aspects of language processing [1], a decision was made to devote this particular study to the more practical aspects, and to emphasize language processing applications in the areas of information retrieval, library and information center processing, text analysis, and translation work.

---

* Dept. of Computer Science, Cornell University, Ithaca, New York

This study was prepared for inclusion in Volume 3 of the Annual Review of Information Science and Technology, and was supported in part by the National Science Foundation under grant GN-495.

Even in this more restricted framework, not all topics are given equal treatment. In particular, some of the problems dealing with the theory and practice of semantic analysis, and their application to the design of question-answering systems are covered more briefly than may be warranted, in part because these questions have been discussed before [1], and in part because the amount of solid, new material in these areas is small. Instead, some of the recent language processing work using on-line consoles, and man-machine interaction is stressed, including in particular interactive systems using mechanized dictionaries and thesauruses, and on-line grammars. Certain practical applications using automatic syntactic analysis procedures are also described in some detail, and an effort is made to assess the effectiveness of these analysis programs.

A number of textbooks and conference reports have appeared within the recent past which may serve for additional reading, and as an introduction to some of the more specialized language processing applications. Bowles [2] covers language processing work in the humanities and social sciences, and Gerbner and Stone [3,4] emphasize more particularly the content analysis aspects inherent in the text inference systems used in the social sciences, where an attempt is made to answer some hypothesis about the content of a text by using, for this purpose, statistical and other language processing methods.

The text by Hays on "computational linguistics" is written largely for linguists who are interested in computer processing. As such, it includes a large number of sample computer programs, and also an introduction to the various syntactic analysis methods now in use. Several books cover language processing with emphasis on applications to automatic documentation. The long awaited text by Borko [6] contains state-of-the-art reports in automatic indexing, automatic extracting and abstracting, and a chapter on question-answering by Simmons, previously published in the journal literature, while Kochen [7] again prints the same Simmons report, as well as chapters dealing with the library of the future. The book by Meadow [8] is only marginally concerned with language processing, since it is written for computer programmers interested in obtaining an introduction to the conventional problems in information retrieval. Finally, the book edited by Schecter [9] contains the proceedings of a well-attended conference in information retrieval, covering a number of interesting language processing projects. Depending on the particular aims and background of the reader, these books may help in a more thorough analysis of the language processing area.

Additional reading matter dealing specifically with the place of language analysis in information retrieval and in the library field is available in survey form [10,11,12], or in the form of citations to the literature in a number of useful bibliographies [13,14,15].

In order properly to put into context the language processing work described in this report, it is useful to be aware of the social and political implications of some of the activities in this area. Indeed, a great deal of organizational work has been devoted in recent years, to the establishment of mechanized information networks, designed to make available to a large user population the combined resources of all components of the network. Thus, design projects exist for library and information networks connecting university facilities [16], for national information networks [17], and for large information centers in certain subject areas designed to cover the field exhaustively [18,19]. Furthermore, it is still occasionally believed to be expedient to point to the existence of well-developed information facilities abroad in an attempt to obtain support for the establishment of large centers and networks in one's own environment. [20]

For obvious reasons, it is not possible to cover the problems connected with the design of mechanized information facilities in the present context. It is, however, important to remember that large information networks are necessarily based on the storage of data files consisting in many cases of written materials and texts. The mechanized storage and dissemination of such materials are likely to raise difficult problems concerning the preservation of ownership rights of the stored texts, and the protection of the private nature of some of the data. One may anticipate that the legal and social aspects of mechanized language processing may take on much greater importance in the near future than they do now, and that they may in time affect also the technical developments.

2. Syntactic and Semantic Theories of Language

A) Automatic Syntactic Analysis

Much of the work in automated language processing is concerned in one way or another with the content analysis of written texts, that is, the problem of transforming the word strings constituting the texts into a formal framework which expresses the information content unambiguously. Content analysis systems are normally based on the recognition of certain key elements — often words in the natural language chosen from a preconstructed list of acceptable items — and on the determination of rules by which these basic elements are combined into larger units. The larger units are of ten word groups with indicators reflecting the relationships between the elements within a group.

Unhappily, while it is relatively easy to isolate individual words in a text, an interpretation of the meaning of the words is much more difficult. Furthermore, no well-defined set of rules is known by which the individual words in the language are combined into meaningful word groups or sentences. Specifically, the correct identification of the meaning of word groups depends at least in part on the proper recognition of syntactic and semantic ambiguities, on the correct interpretation of homographs, on the recognition of semantic equivalences, on the detection of word-word relations, and on a general awareness of the background and environment of a given utterance.

While syntactic analysis procedures cannot be used to solve the complete content identification problem, the syntactic properties of words are important in recognizing certain relationships which exist between words within a sentence, reflected, for example, by noun phrase combinations, prepositional phrases, adverbial phrases, and simple subject-verb-object groupings. It is for this reason that syntactic analysis procedures are normally included in many automatic content analysis systems.

Most automatic syntactic analysis systems are based on construction rules, or grammars, known as <u>phrase structure grammars</u>, in which a sentence in the natural language is regarded as consisting of a set of juxtaposed or nested phrases. A phrase structure grammar is normally defined by a set of rewrite rules, and the derivation of a given sentence producible by a phrase structure grammar is specified by citing the rewrite rules used in generating it, as well as the order in which the rules are applied. The derivation of a sentence in a given grammar can be represented by a tree called a phrase marker or structural description. A typical phrase marker is shown in Fig. 1 for the sentence "John is easy to please". [21,22]

It may be seen from Fig. 1, that the rewrite rules are used first to expand the symbol "sentence" into a string consisting of the symbol "NP" (noun phrase) followed by the symbol "VP" (verb phrase), or

$$(\text{sentence}) \rightarrow (\text{NP}),\ (\text{VP}) \quad .$$

Thereafter, the symbol "VP" is expanded further into "copula" followed by "predicate", or

$$(\text{VP}) \rightarrow (\text{copula}),\ (\text{predicate}),$$

and so on, down the tree, until the various intermediate symbols are replaced by the final words of the sentence being constructed. The "phrases" actually recognized in the sample sentence can be marked off by pairs of brackets as follows:

(easy to please)
[is (easy to please)]
{John [is (easy to please)]} .

Phrase structure grammars suffer from various well-known disadvantages which diminish their potential usefulness as part of an automatic content analysis system. In the first place, a large number of different derivations are often produced for a given sentence, and no methods are available for choosing the one correct analysis which may apply in a given instance. Furthermore, some of the resulting analyses, while grammatically correct, may be semantically unacceptable. Thus, for a sentence such as "they are flying planes", a typical phrase structure grammar may generate three different analyses. The first accepts "flying planes" as a complement of the predicate verb "are", so that "they" refers to "planes". In the second, "they" refers to people, and "flying" is a present participle of the transitive verb. The third analysis is semantically absurd, and reflects the structure of a sentence such as "the facts are smoking kills", which is not semantically absurd; in that case, "flying" is identified as the gerund of a transitive verb which acts as subject of a complement clause, and "planes" is accepted as the complete intransitive verb which acts as the predicate of the subclause. The same three analyses would be obtained for the sentence "the facts are smoking kills" but only the third one would be semantically correct for that sentence. [23]

Another drawback of phrase structure grammars relates to the fact that such a grammar reflects only the surface structure of each sentence which accounts for the phonetic representation, but not necessarily for the semantic interpretation. Thus, in the sample sentence of Fig. 1, "John" is taken as the subject of the verb "is", and "easy to please" is part of a complementing verb phrase. A completely different theory of grammar, known as a transformational grammar has, therefore, been developed in recent years, designed to take into account not only the surface but also the deep structure of a sentence which accounts for the semantic interpretation. [24,25] A sample deep structure for the sentence "John is easy to please" is shown in Fig. 2.

It may be seen that in the deep structure, "John" is correctly identified as the object of the verb "please", and that the subject of that verb is some unspecified person designated as "someone". Furthermore, the subject of "is easy" is not "John", but the noun phrase given on the left-hand side of the tree structure. On the other hand, the fact that "John" is the subject of "is easy" in the surface structure cannot contribute to the proper semantic interpretation of the sentence. A comparison of Figs. 1 and 2 exhibits the limited utility of phrase structure grammars in specifying both syntactic and semantic interpretations.

A transformational grammar consists of two main components, known as base component, and transformational component, respectively. The base component generates for each sentence a set of terminal strings with an associated structural description of the type shown in Fig. 2, and known as generalized phrase markers. The transformational component, on the other hand, maps phrase markers into other phrase markers by a sequence of simple transformations including substitutions, deletions, and adjunctions. For example, the transformational component can be used to turn a declarative sentence into an interrogative one, and more generally to accept as input a generalized phrase-marker and to produce from it a derived phrase marker.

The recognition procedure for a transformational grammar, that is, the process used to construct the generalized phrase marker for each given sentence is much more complex than that for a phrase-structure grammar, largely because the grammar itself is not expressible by simple kinds of rewriting rules. Since a transformational grammar is capable of mapping deep structures into surface structures, most proposed recognition procedures make use of reverse transformations which start with a derived phrase marker and attempt to generate from this a generalized phrase-marker. Specifically, a phrase structure grammar is used first to generate one or more derived phrase-markers for each sentence. A reverse transformational component is then used to generate from the derived phrase-markers a set of final phrase markers displaying the deep structure. Those deep structures are then identified as correct semantic interpretations of a given sentence for which the corresponding generalized phrase-marker can be mapped by the (forward) transformational grammar into a phrase-marker which is exactly the same as the derived phrase marker generated by the transformational grammar. Since little experience is available concerning the operational performance of transformational components, it is not known how easily such matching structures are, in fact, obtainable.

The research performed over the last few years in applying syntactic theories to language processing problems falls into several categories. Simplified phrase structure grammars have been used extensively in an attempt to generate by relatively inexpensive means at least some of the major phrase groupings. This work is further described in section 4. More refined syntactic methods using transformational grammars have also been initiated. However, since transformational grammars are more difficult to implement, efforts have been concerned either with the abstract characterization of tree transformations [26], or else with the design of semi-automatic analysis systems, in which the user is given a chance to speed up the analysis by making appropriate suggestions during the course of the analysis [27]. These semi-automatic efforts are further outlined in section 3.

B) Automatic Semantic Analysis

It is clear from what has already been said, that no known automatic syntactic analysis method is available for generating a unique semantic interpretation for each given well-formed English sentence. At best, certain word groupings (phrases) can be isolated within the boundaries of individual sentences, but the relationship between phrase components, and between individual phrases remains largely unknown. For practical purposes, it is, in fact, necessary to rely on phrase-structure grammars for the analysis of natural language input, rather than on the more refined transformational grammars since the latter are still undefined in part and expensive to implement. Under the circumstances, many analyses are produced for each sentence, some of which may be semantically absurd.

Even if it were assumed that absurd analyses could somehow be discarded, and that (transformational) grammars were implemented which would assign the same analysis to many derived structures (such as "John is easy to please", and "it is easy for someone to please John"), there would still exist no solution to the problem of choosing the appropriate meaning for words which can normally carry many different meanings (even within the same contexts]. Thus, in the sentence "they are flying planes", it is not possible to tell by any known formal analysis procedure whether any given occurrence of "they" refers to planes or to people.

Several models of semantic analysis have been proposed, based on a dictionary of entries furnished with appropriate semantic markers, and on a transformational grammar for the recognition of syntactic structure. However, no model so far proposed appears to be fully adequate for the task at hand. Simmons and Burger list the following five components of an ideal natural language processor [28]:

a) a syntactic analyzer used to identify structural relations;

b) a semantic analysis to transform the syntactic output into unambiguous entities in some formal language;

c) a logical (cognitive) structure of objects and relations that represent the meanings of the entities as perceived by humans (normally specified by a semantic dictionary);

d) an inference procedure for the recognition of distinct syntactic structures with equivalent meanings; and

e) a syntactic-semantic generation system for producing English statements from certain given formal structures.

Since it is convenient not to deal with the full magnitude of the semantic analysis problem all at once, most of the more theoretical work in semantics is, in fact, restricted to only one or another of these tasks. Thus, some effort is devoted to the construction of semantic dictionaries incorporating appropriate definitions useful in a semantic characterization of each entry [29,30,31]. The problem of word disambiguation is also considered by a number of authors who try to use contextual rules in order to uniquely determine the particular meaning of each occurrence of a text element [32,33,34]. Finally, some attention is given to a characterization of discourse structure where a complete text is analyzed as an entity, rather than on a sentence-by-sentence basis, and the various linking devices used in the language are identified, as well as the referents to other parts of the text [35,36].

Many researchers argue that, in light of the difficulties which emerge in any serious attempt at a complete automatic language analysis, the emphasis should be placed on the manual simulation of perfect analysis systems, or, alternatively, on the use of imperfect, partly manual systems, designed to solve some interesting, practical language analysis problem in a restricted environment. The latter approach is the one generally taken by people interested in the so-called question-answering problem, where automatic, or semi-automatic methods are used to answer questions submitted by a user population.

A number of models exist for question-answering, including some developed within the last year or two [37,38,39]. A simplified diagram of the model by Woods [37] is reproduced in Fig. 3. It may be seen that the system consists of three parts, including a syntactic analyzer which generates the (deep) structure of an input query, a semantic interpreter used to obtain the meaning of the query in terms of certain formal entities specified by a stored data base, and finally, a retriever which matches the semantic structures resulting from the semantic interpretation of the query with the structures recognized in the data base, and constructs, or computes, suitable answers.

The stored data base in the model by Woods consists of objects, as well as functions which map sets of objects into other objects, relations which replace some of the verbs and prepositional modifiers in the natural language, and prepositions which are instances of particular relations between objects. The semantic interpretation is then effectively a translation of the syntactic structure of a sentence into an expression in the formal query language representing the meaning of the sentence in terms of predicates, functions, and commands understood by the system. The meaning of each predicate, function, or command is defined by a programmed subroutine that generates the truth value of the given predicates, or computes the functional values for the arguments which may be present in the given expression.

It should be noted that the various parts of the semantic language processor previously outlined by Simmons are all present in the question-answering model under discussion, although the restricted nature of the data base -limited to objects, functions, and relations dealing with airline time tables -renders a sophisticated inference procedure for the recognition of equivalent semantic structures unnecessary. Similarly, the process of constructing answers to queries is also simplified because of the special environment within which the system operates.

In the system by Woods, the semantic interpretation is specified by a set of replacement rules, similar in spirit to the rewrite rules used in a phrase structure grammar. Specifically, for each given input pattern, represented by fragments of deep phrase structure diagrams with specific entities attached to the bottom (terminal) nodes of the tree fragments, a certain action is called for to specify the meaning of the input pattern.

Thus, if the input statement is "AA-57 flies from Boston to Chicago" as shown in Fig. 4(c), and the structure of the various phrases is specified respectively by tree fragment G1 for "AA-57 flies", and tree fragment G3 for both "from Boston", and "to Chicago", then the semantic rule reproduced in Fig. 4(a) replaces the input by the formal statement

CONNECT (AA-57, Boston, Chicago).

The predicate "CONNECT" takes three arguments and is defined in the system by an appropriate subroutine which determines its truth or falsity for any three given arguments. The system can answer yes/no questions, as well as a variety of other requests for which an answer can normally be looked up in an Official Airline Guide [40].

The input statements are thus transformed into sets of nested functions, each of which must be evaluated by reference to the data base before an answer can be generated. For example, a query such as

"how many flights go from Boston to Chicago"

is transformed by the syntactic analyzer and semantic interpreter into the formal statement

LIST(NUMBER(XI/FLIGHT: CONNECT (XI, Boston, Chicago))).

that is, list the number of XI's, such that XI is a flight, and the CONNECT predicate is true.

The question answerer by Woods is typical of the more advanced work in this area, and demonstrates the complexity of the rules and size of the grammar actually needed to handle queries in a restricted environment. Whether such systems can be extended to other data bases and to different environments remains to be seen. One might anticipate that the predicate set, as well as the set of functions and relations would have to become very large, and that the semantic rules might start to contradict each other, or might no longer be applicable unambiguously. In any case, much additional work along those lines will be required over the next few years before the full potential of experimental systems of this type becomes known.

Whereas Woods [37] and Kuroda [41] generate nested function calls as output of the semantic interpreter, other workers transform the English input into standard language "kernels" [42,43], or alternatively, they operate directly with the deep structures resulting from the transformational syntactic analysis [44]. A large number of researchers also use ad hoc contextual rules to resolve ambiguities during the semantic analysis [33,45], or else they unashamedly use a simplified input language which is unambiguous to begin with, thus avoiding the complicated and time consuming analysis procedures previously outlined.

A number of workers dealing with document retrieval systems, instead of question answering, believe that a complete semantic interpretation of each input text and search request is necessary in that case also, if the system is to operate effectively [49,50]. Others, however, maintain that a unique interpretation of texts is not wanted in document retrieval [50]. An attempt is made to resolve these questions in section 4, when text processing applications are covered in more detail.

3. On-Line Text Processing

A) General Linguistic Analysis

Probably the most noticeable development in language processing over the last few years is the increasing tendency to use on-line, machine-aided techniques instead of completely automatic methods for linguistic analysis. Specifically, a research worker introduces programs and data by sitting at a slow-speed computer terminal — typically a computer controlled typewriter unit. Such a terminal then permits the user directly to monitor the progress of the operations, and to help in the analysis by choosing appropriate alternatives for multiple analyses, or by rejecting erroneous steps.

The trend toward semi-automatic methods is the result of the development of time-shared computer organizations, which permit a multiplicity of users to share computing facilities on a cooperative basis, and also of the realization that many language analysis tasks are not about to be solved satisfactorily by fully-automatic means. The well-known ALPAC report [51], which suggests that fully-automatic machine translation should not be regarded as a realistic goal for the immediate future, may have provided an additional incentive for the move to on-line, interactive processing.

A great deal of work has thus been devoted to the generation and manipulation of mechanized dictionaries and thesauruses, and to language analysis methods using on-line grammars. These efforts are described in the next two subsections. In addition, a large variety of linguistic manipulations are also being done semi-automatically with appropriate user interaction, including, for example, the well-known text editing application [52,53]. Sutherland [54] is using a light pen pointing device attached to an input-output display system to introduce into the computer complex linguistic commands. In the same spirit, an on-line console is being used as a transcription device for the introductions into a computer of Chinese characters [55]. Specifically, a cathode ray tube display serves first for the generation and storage of the various characters and radicals. Thereafter, Chinese text can be introduced without special encoding by having an operator point to the appropriate character wanted at any given time, assuming, of course, that the character set is properly displayed.

In such a system, the operator is instrumental in speeding up the search for the character needed at each moment. Such a semi-automatic text input system can be attached to a mechanized text manipulation or translation system which later furnishes output in the appropriate form.

On-line consoles may also be used experimentally for linguistic field work to determine the characteristics of some unknown target language under study [56]. Specifically, queries are transmitted (in English) to a bilingual informant sitting at a console; the informant then responds by giving appropriate responses (in English) about his language; the system then makes inferences about the target language and formulates further queries which are again transmitted to the informant. It is hoped that such a system may eventually be sufficiently powerful to be used in constructing morphological or syntactic rules for unknown languages.

Display systems are also used for more general file processing tasks in a tutorial mode by using the system to display a variety of options which may be available at any given time, and letting the user respond appropriately in each case, thus initiating further processing leading to additional options submitted to the user [57,58,59].

Work is also proceeding in the automatic generation of correct English output responses supplied by the computer in answer to appropriate input stimuli received from a user sitting at a console [60]. This work is intended eventually to permit smooth interaction with computers, using relatively free English input.

B) Automatic Thesaurus Processing

Stored dictionaries and tables have been used for years in order to obtain functional correspondents for included dictionary items, for example, a translation into a foreign language of each stored word, or a storage address for the record corresponding to a given input key. The large variety of new applications using dictionary operations, and the use of interactive methods to control the display of appropriately chosen dictionary excerpts is, however, relatively new.

In a well documented study, Galli and Yamada [61] explain how an English-English dictionary is used to classify English words into one of seven morphological groups; the classification serves as part of an automatic program for vocabulary control, and includes features for hyphenation, correction of spelling errors, and proofreading. Mechanized special purpose dictionaries of various kinds are used for vocabulary control also in many other technical fields [62].

A mechanized dictionary also forms the basis of a study which may eventually lead to a complete morphological and semantic analysis of the language [63]. Specifically, a complete issue of Webster's English dictiona is key punched, and is used for the systematic generation of English word stems, and for a characterization of the minimal set of semantic primitives which are needed to describe the semantic component of the language. It is hoped, in particular, that a much smaller set than the 150,000 different word senses now used by Webster to characterize the English language may prove adequate.

Mechanized multi-lingual dictionaries are also used as components in semi-automatic machine translation systems [64,65] in such a way that various dictionary entries are displayed during the translation activities, while leaving it up to the user to guide the system by making appropriate choices at critical points — for example, by resolving ambiguities. It is hoped in this fashion to produce a translation which is sufficiently accurat to require no post editing. Margaret Masterman discusses the potential of such a computer-aided translation, and points out that a mechanized dictionary, even when used semi-automatically will likely lead to errors because of the tendency to look up one word at a time, and thus, to translate on a word-by-word basis instead of globally [65].

Considerable effort has also gone into studies dealing with the automatic construction of dictionaries [66,67]. Such work is designed to overcome the time lag normally inherent in any manual dictionary construction. It would be nice to be able to report that fully-automatic dictionary generation can be expected soon for most applications. Unfortunately, this is not likely to be the case for some time to come, and reliance must, for the most part, still be placed on a variety of manually prepared dictionaries [68,69,70].

C) On-Line Grammar Processing

An increasing amount of work is being done with syntactic systems used in an interactive mode. Typically, a grammar is first tested in a tutorial mode by displaying the rules one at a time, and demonstrating their application to the derivation of surface and/or deep structures for various input sentences. The user is then permitted to amend or reject certain rules under certain circumstances, thereby refining the grammar as the work proceeds.

In recent times, most efforts of this type have been based on transformational grammars whose structure is not known as well as the more familiar phrase structure grammars. The user interaction is then designed to let the user follow the derivational history of each sentence, and to test the applicability of each rule in a step-by-step fashion [71,72,73,74]

Once the grammar is sufficiently tested, it can then be applied to certain language processing tasks — normally the generation of deep structures, or language kernels, to be used later in a question-answering or data retrieval system [75,76,77].

In the present state of research, consisting mostly of relatively small-scale attempts to use on-line grammars for restricted purposes, it is difficult to say whether the on-line mode really produces the kind of environment which may eventually result in a substantial simplification of the language analysis process. Much work remains to be done before such questions can be answered with assurance; in the meantime, one may look forward to additional experiments in the construction and testing of increasingly more powerful grammars using on-line devices.

4. Syntactic Processing for Document Retrieval

A number of research efforts were covered in previous sections whose aim is the complete semantic analysis of a given input text, including the resolution of all applicable ambiguities. While such efforts are, of course fully justifiable, both from a theoretical and from a practical viewpoint, there exists also an alternative approach which consists in using a simplified type of analysis designed to extract as much knowledge from a text as could be obtained rapidly and inexpensively. No attempt is then made at complete disambiguation; instead, the attitude is to make do with what is easily at hand, and to see how far one gets in actually applying such techniques to operating text processing — chiefly retrieval — systems. Syntactic procedures of this type are covered in the present section, while statistical processes are examined in section 5.

Typical of the kind of effort under discussion is a syntactic analysis system used by Hillman and his group [78,79,80], in order to generate noun phrases used for document and query identification. The process is based on a small, easily constructible dictionary, containing only a restricted set of special function words such as prepositions, conjunctions, auxiliary verbs, and some suffixes. Each text word is looked up in this dictionary, and items actually found are assigned their appropriate part of speech. Other items not included in the dictionary are marked as unknown. A computer program is then used to determine the syntactic markers of all unknown items by analyzing the pattern of the surrounding function words. After syntactic markers have been assigned to all text items, a series of reduction steps is used to identify noun phrases, verb phrases, and prepositional phrases; these intermediate phrases are then used in turn to specify the principal noun phrases and their qualifiers which serve for purposes of phrase identification.

An example of such an analysis is shown in Fig. 5, for a document abstract in the area of documentation. Each text item is followed by the syntactic marker assigned during the analysis, and the noun phrases are identified by pairs of NP markers (the brackets have been added for convenience in reading, and are not part of the normal output). It may be seen from the output of Fig. 5 that quite a few of the noun phrases are identified correctly. Others, such as "each specialized center" are not recognized, because the adjective "specialized" is mistakenly identified as a past participle; furthermore, "through" in the last sentence of the text is recognized as a noun. In many cases, such a simplified analysis will also be unable to recognize the main verb accurately, and sentences with subordinate clauses are often badly analyzed since the analysis scheme assumes that such sentences are always made up of a complete main clause followed by the complete subordinate clause.

Simplified syntactic analysis methods based mainly on the presence or absence of function words have previously been used for a variety of purposes [81,82,83]. They are attractive because of the relatively small apparatus needed in their operation. For information retrieval purposes, such methods are likely to be deficient, because a document identification consisting of complete syntactic phrases instead of simpler content identifiers such as single terms, may represent an over-specification of the content which may result in a deficient recall performance (that is, some documents not identified by the chosen phrases may not be retrievable even though pertinent). In addition, the generation of false phrases may also depress precision by helping to retrieve items which are, in fact, not wanted.

Hillman expects to overcome these problems by using an interactive system which supplies new related phrases to be used as part of the search specification, in addition to the phrases originally chosen by the user in his query specification. In fact, all phrases present in the system are arranged into groups, called genera, as a function of the co-occurrence characteristics of these phrases in the documents of the collection. An on-line display system, similar to those previously discussed, is then used to add to the original queries associated phrases located in the same genus as those originally used. Documents containing phrases also included in the search requests are then retrieved in answer to the queries. Based on th retrieval evaluation work with the SMART system [84], it is likely that much simpler document analysis methods, not based on syntax, would work at least equally well.

Such a conclusion may be confirmed by a somewhat related experiment by Melton in which an attempt was made to compare the effectiveness for retrieval purposes of syntactic phrases with that of unstructured content words included in documents and search requests [85]. The results of that particular study are difficult to interpret, because the relevance judgments used in the evaluation study were based on document representations which were different from those used in the search operation, and also because retrieval criterion apparently considered the presence of only a single matching phrase in both document and search request to be sufficient for retrieval. The Melton study does, however, appear to confirm the previously published results of the SMART work which indicate that syntactic phrases derived from document sentences represent an indexing tool which is normally too specific for document retrieval purposes [86].

The reasons for this unexpected result may be both with the relatively inadequate syntactic tools presently used in a sentence-by-sentence mode for purposes of content analysis, and also with the availability of rather refined statistical and dictionary analysis procedures, which in the aggregate seem to outperform the syntactic processes in a document retrieval environment.

A number of studies have been initiated recently with the intention of upgrading the retrieval performance obtainable with syntactic phrase matching methods by using more refined syntactic analysis procedures, rather than by replacing the syntax with alternative statistical processes. The well-known Syntol work is a case in point, where a dictionary of syntactic templates (grilles syntaxiques) is used to characterize a text by an abstract graph whose nodes represent unambiguous Syntol concepts, and whose branches represent specified relations between corresponding pairs of concepts [87,88,89,90].

A great deal of work is devoted at present to programs designed to carry out an automatic oncstruction of the graphs, as well as a complete disambiguation of all its components. No evaluation data are, however, available which could be used to compare the effectiveness of this type of analysis with other alternative procedures.

Another recent effort in a similar direction is based on a syntactic analysis process known as string analysis which produces phrases similar to those illustrated in Fig. 6 for the document previously used as an example in Fig. 5 [91,92,93]. The output in Fig. 6 can be read by successively replacing the numbers on the right side of the expressions by their definitions on the left. Thus, the phrase "each specialized center" is correctly analyzed by the string analysis of Fig. 6 as follows:

(from 4 center or 5)
(from (each specialized) center or 5)
(from (each specialized) center or through 6 center)
(from (each specialized) center or through (the AF materials) center).

The string analysis retrieval system is again based on a matching of phrases contained in documents and search requests. It also uses, during the analysis, a semantic dictionary containing synonyms, a variety of term relations, definitions, tables of data, and specifications for various types of operators. Since the proposed retrieval system is not implemented, its effectiveness cannot be assessed. The precision is likely to be high, becaus few erroneous phrases will normally be generated. On the other hand, the recall may be low, compared with that obtainable by other less powerful analysis methods.

One other recent study deals with the automatic generation of phrase for purposes of content identification [94]. Jones and Giuliano use mostly statistical criteria for selecting the so-called "content bearing units", including co-occurrence statistics for the various components of a phrase, as well as individual relative and absolute frequency statistics of the components. The statistical model is then supplemented by a minimum of grammatical information consisting of probability coefficients for the syntactic class of each component of a phrase.

The techniques for choosing the content bearing units is interesting, and it is claimed that it leads to high precision searches. However, since only 25 to 45 percent of the actually valid word pairs are said to be selected, the method may again suffer from a low recall performance. In any case, like the other previously described systems, the process using content bearing units also remains to be tested in a retrieval environment.

5. Statistical Processing for Document Retrieval

Statistical procedures, based partly on word frequency characteristics, and partly on co-occurrence criterion between words in sentences, or paragraphs, or complete documents, have been used for over a decade for text analysis purposes. Within the past year, statistical methods have thus been incorporated into systems for the automatic recognition of proper names [95], for the proper segmentation of sentences [96], for the study of English affixes [97], for the analysis of style in written texts [98], and for the automatic identification of synonyms and antonyms [34].

A statistical technique which in recent years has become almost universal for a variety of applications is that of computing statistical associations of a given set of entities. Typically, each item originally available may be specified by a given property set; by comparing the property sets for each pair of items, it becomes possible to obtain an association coefficient between the corresponding items, based on the similarity of their property sets. Items whose association coefficient is sufficiently high are then taken to be associated or related.

Associative procedures are used to expand the term vectors of individual documents by adding new associated terms to the ones originally available. They can also serve to expand document groups by adding to each given group new documents, found to be associated with the original one: In dictionary construction or utilization, associated terms can be added to the originally available dictionary classes so as to provide an expanded vocabulary.

Associative techniques are also important in automatic systems designed to identify text words likely to be useful as content indicators [99,100,101]. In the well-known experiment by Dennis [102], the ratio of raw occurrences of a word to the reciprocal of the coefficient of variation of its within document frequency is used as a means to identify content word (if the latter coefficient is large, the assumption is that a content word has been identified). The original content words together with significantl associated words are then used as index terms for the identification of docu ments, and for search purposes. Dennis finds that the addition of associate terms significantly improves search effectiveness.

A variety of association coefficients have been proposed in the literature [103]. Ultimately, they all use co-occurrence characteristics as an indication of significant association. That is, if two terms co-occur sufficiently often in the sentences of a document, or in the documents of a collection, or if two documents are identified by a sufficient number of joint content indicators, the corresponding items are assumed to be associat and are included in the same class.

In the associative retrieval model proposed by Giuliano and Jones, the additional assumption is made that documents and index terms are linear! related [104,105]. Furthermore, the thought is expressed that first order associations between words represent so-called "contiguity" relations, whereas second order associations, corresponding to items which have similar contexts, represent synonymy relations. Recent evaluation results for such associative retrieval systems suggest that the associated terms do, in fact, improve retrieval effectiveness, compared with systems which use only raw words originally present in documents [106]. However, well constructed thesauruses appear to provide better language normalization than the statistically determined term associations; furthermore, the conjectures concerning the role of associated terms as items denoting synonymy or contiguity appear not to be borne out in practice [106].

Experiments using term associations and term profiles as a means for expanding queries or document identifications may be expected to continue for some time to come, particularly in view of the increasing use of interactive techniques for retrieval purposes [107,108].

One of the most interesting new developments in the area of statistical language analysis has recently taken place in the area of automatic classification using clustering or clumping techniques [109,110,111]. Such programs normally start with a matrix of order n^2 giving the association measure for each possible pair of n available items. Potential clumps are formed by matching each item against each other, and items are then shifted from one clump to another until such time as the "cohesion" of the clumps can no longer increase.

The clumping procedure is normally impossible to carry out for all but the smallest data sets because the n^2 process which results from matching each item with each other becomes exorbitantly time-consuming. A new procedure proposed recently replaces the n^2 process by one of order $n \log n$ which promises to be more manageable [112,113]. This new clustering procedure starts with an attribute list for each item to be clustered, from which are constructed group profiles for groups of items initially assigned arbitrarily. Items are then shifted from one group to another until equilib rium is attained. The procedure is guaranteed to converge under certain circumstances [113], but has not apparently been used under operational conditions.

6. Language Processing Systems

In the available space, it is not possible to provide more than a cursory view of the present state of operational automatic text processing systems. Because of the impact of the ALPAC report already referred to [51,114], practical work in fully-automatic machine translation seems more or less at a standstill. Instead, the machine translation work consists of theoretical investigations [115,116], or alternatively, of experiments in machine-aided translation in which an exhaustive analysis system is replaced by the presence of a human intermediary [117]. Some work is also being done on the evaluation of the accuracy and effectiveness of various translation systems [118].

Interest continues to grow in automatic systems for the selective dissemination of information (SDI) which are designed to bring to the users' attention, citations to the literature believed to be of interest to each recipient [119,120,121]. Such services compare stored profiles describing user interests with stored document identifications in order to match available documents with potential recipients. The designers of such systems claim high precision performance, but the recall figure is not normally given (relevant items might be included in the file which are never distributed to users who would, in fact, find them of interest); furthermore, SDI systems suffer from the fact that some users may not necessarily want to receive the notifications which are automatically provided by the system, even though the corresponding items may, in fact, be perfectly relevant.

On-line document retrieval systems using man-machine communication to help in the retrieval activities were mentioned earlier. Some of these experimental systems are well known and have been extensively described in the literature [122,123,124]. Others are presently under construction such as a proposed on-line system at the University of Pennsylvania [125,126,127, 128]. User interaction is also used in the SMART system in conjunction with iterative search methods which allow the user to furnish information about previous search results in an attempt to improve performance in subsequent search iterations [129,130].

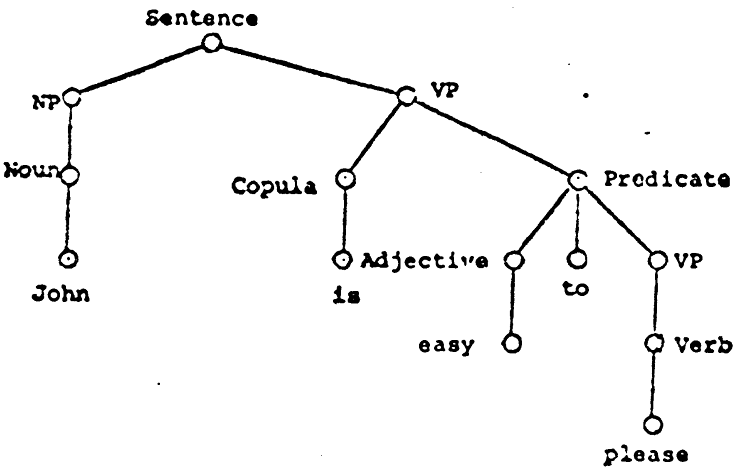
Finally, a variety of experimental question-answering systems are in operation, including one in which syntax and semantics are interwoven in an interesting way [131], one able to make inferences and to answer questions about simple pictures [132], and one which incorporates an on-line learning system designed to make the user aware of the available data bases and of the various facilities which may be called upon during the retrieval activities [133,134].

For practical purposes, it may be expected that for some time to come, most question-answering systems will deal with strictly circumscribed data bases expressed in a substantially unambiguous formal language; an example of such a system with a practical orientation is the system by Levien and Maron based on a store of triplets of the form "entity i — relation j — entity k", expressing a specified relational type between a given pair of entities [135]. The general critique of question-answering systems by Kasher [136] may be used as an introduction to the field which may still have validity today.

Most investigators are aware of the increasing need for an evaluation of automatic text processing systems so that the proper understanding may be gained before such systems are implemented under operational conditions. In fact, a great deal of effort is being devoted to various evaluation experiments. Several projects are engaged in a comparison of a variety of different automatic text processing methods [84,137]. Others base their evaluation on cost criteria [138], and others still are setting up large test environments under controlled conditions [139,140,141,142].

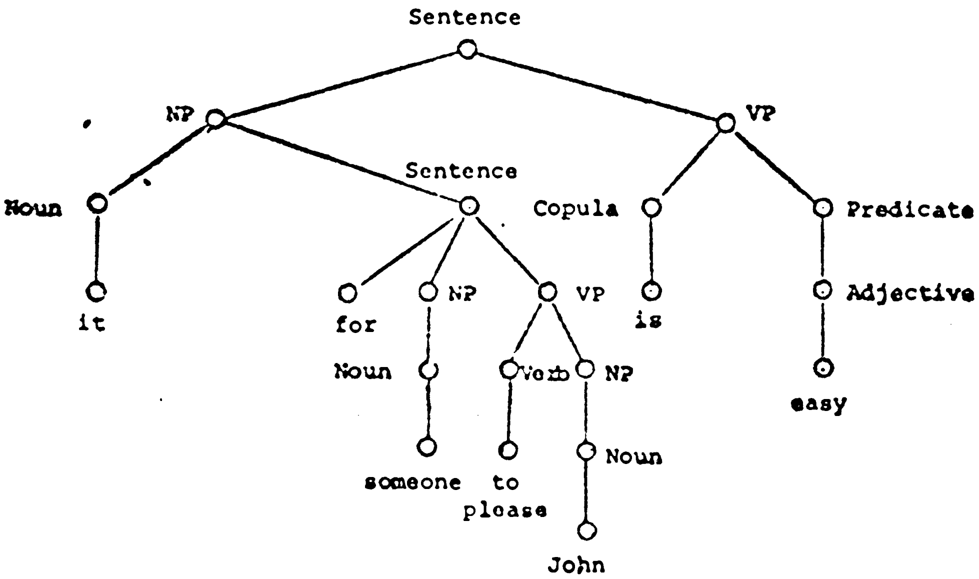
It is easy to find, in the literature, opinions of various sorts expressing misgivings about the effectiveness of automatic language processing techniques; a typical example may be the recent statement by Sharp that "it is disappointing that the early and ingenious ideas of H. P. Luhn for matching terms against text on a purely statistical basis have proved to be abortive". [11].

While a complete evaluation of all automatic text processing work remains to be done, it is already clear that many of the negative statements cannot be supported by experimental evidence without qualifications. Obviously, among the many possible techniques, not all prove equally effective for all applications. What proves useful in document retrieval may not avail in machine translation, and what helps the user interested in high precision may not do for the customer who needs high recall. As more experience is gained with automatic language processing techniques, it may eventually be possible to use exactly the right degree of sophistication necessary to provide effective output in each particular environment. In any case, one may expect that increasing use will be made of refined statistical and syntactic analysis procedures, and that the semantic analysis will increasingly be carried out in an interactive mode with the help of the user. It is hoped that the present study may serve as a preview of things yet to come in the language processing area.



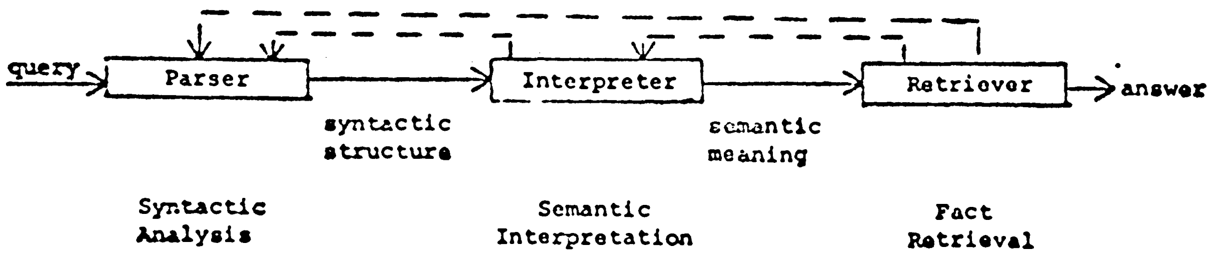
Phrase Marker (Surface Structure)

Fig. 1



Sample Deep Structure

Fig. 2



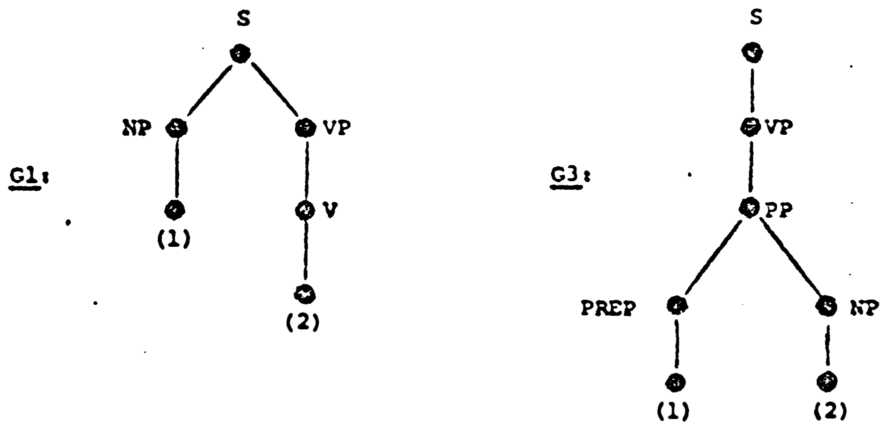
Model of Question-Answering System

Fig. 3

```
1.  (G1:  FLIGHT ((1)) and (2) = fly)  and
2.  (G3:  (1) = from and PLACE ((2)))  and
3.  (G3:  (1) = to and PLACE ((2)))

      ==> CONNECT (1-1, 2-2, 3-2)
```

a) Typical Semantic Rule



b) Tree Fragments Used for Semantic Rule of part (a)

AA-57 flies	from Boston	to Chicago
matches first application of tree G1	matches first application of tree G3	matches second application of tree G3

c) Sample Input Sentence Corresponding to Semantic Rule of part (a)

Semantic Interpretation in System by Woods

Fig. 4

D12 Title: NP{The A materials A information A network N}NP

NP{The A elements N}NP PS{of P NP a A network N NP}PS PS{of P NP information A centers N NP}PS are V described V.

NP{The A various A files N}NP and Cl NP{retrieval A procedures N}NP PS{at P NP each A data A center N NP}PS are V outlined V.

NP{Information N}NP can V be V retrieved V directly B PS from P NP{each N}NP specialized VD NP{center N}NP or Cl NP{through N}NP NP{the A Air-Force A materials A information A center N}NP.

Output of Hillman's Syntactic Analyzer for Document D12

Fig. 5

D12. Title:

0.	noun	=	1 network
1.	art adj	=	the materials information
0.	sentence	=	7
7.	assertion	=	1 elements 3 are 2
1.	art	=	the
2.	passive	=	described
3.	prep	=	of 4 network 5
4.	art	=	a
5.	prep	=	of 6 centers
6.	adj	=	information
0.	sentence	=	7
7.	assertion	=	1 files and 2 are 3
1.	art adj	=	the various
2.	noun	=	4 procedures 5
4.	adj	=	retrieval
5.	prep	=	at 6 center
6.	adj	=	each data
3.	passive	=	outlined
0.	sentence	=	7
7.	assertion	=	information can be 1
1.	passive	=	retrieved 2 3
2.	adv	=	directly
3.	prep	=	from 4 center or 5
4.	adj	=	each specialized
5.	prep	=	through 6 center
6.	art adj	=	the Air Force materials information

String Analysis Output for Document D12 of Fig. 5

Fig. 6

References

[1] D. G. Bobrow, J. B. Fraser and M. R. Quillian, Automated Language Processing, in Annual Review of Information Science and Technology, Vol. 2, 1967, C. Cuadra, editor.

[2] E. A. Bowles (editor), Computers in Humanistic Research Readings and Perspectives, Prentice Hall, 1967.

[3] G. Gerbner (editor), Reprints of the National Conference on Content Analysis, University of Pennsylvania, November 1967.

[4] P. J. Stone, D. C. Dunphy, M. S. Smith, and D. M. Ogilvie, The General Inquirer: A Computer Approach to Content Analysis, M.I.T. Press, Cambridge, 1966, 651 pages.

[5] D. G. Hays, Introduction to Computational Linguistics, American Elsevier Publishing Co., 1967.

[6] H. Borko (editor), Automated Language Processing — The State of the Art, John Wiley and Sons, 1967.

[7] M. Kochen, The Growth of Knowledge, John Wiley and Sons, 1967.

[8] C. T. Meadow, The Analysis of Information Systems, A Programmer's Introduction to Information Retrieval, John Wiley and Sons, 1967.

[9] G. Schecter (editor), Information Retrieval — A Critical View, Thompson Book Co., Washington, 1967. (Proceedings of the 3rd Annual Colloquium on Information Retrieval, Philadelphia, 1966).

[10] M. Coyaud, Les Methodes Modernes de Documentation, Report P.1/SAD, Paris, December 1966.

[11] J. R. Sharp, Content Analysis, Specification and Control, in Annual Review of Information Science and Technology, Vol. 2, 1967, C. Cuadra, editor.

[12] B. W. Adkinson and C. M. Stearns, Libraries and Machines — A Review, American Documentation, Vol. 18, July 1967, pp. 121-124.

[13] D. G. Hays, B. H. Dostert and M. L. Rapp, Computational Linguistic Bibliography 1966, Rand Memorandum RM 5345-PR, April 1967.

[14] D. G. Hays, B. H. Dostert, and M. L. Rapp, Annotated Bibliography of Rand Publications in Computer Linguistics, Rand Memorandum RM-3894-3, June 1967.

References
(contd)

[15] F. Heeland, A Bibliography on Information Science and Technology for 1967, System Development Corporation Report TM(L)-3553, Santa Monica, California, 4 parts: July 20, 1967, October 1, 1967, December 1, 1967, January 15, 1968.

[16] G. W. Brown, J. G. Miller and T. A. Keenan, EDUNET — Report of the Summer Study on Information Networks, John Wiley and Sons, New York, 1967.

[17] L. F. Carter, National Document Handling Systems in Science and Technology, Science, Vol. 154, December 9, 1966, pp. 1299-1304.

[18] F. A. Tate, Handling Chemical Compounds in Information Systems, in Annual Review of Information Science and Technology, C. Cuadra, editor, Interscience, Vol. 2, New York, 1967.

[19] LITE, New Total Text Information Storage and Retrieval System, Scientific Information Notes, Vol. 9, June 1967, p. 3.

[20] A. I. Mikhailov, Development and State of the Art of the Scientific and Technical Information System in the U.S.S.R., 33rd Conference of FID, Tokyo, September 1967.

[21] D. G. Bobrow, Syntactic Theories in Computer Implementations, in Automated Language Processing, Borko (editor), John Wiley and Sons, 1967.

[22] B. Vauquois, G. Veillon, J. Veyrunes, Syntax and Interpretation, Mechanical Translation, Vol. 9, June 1966, pp. 44-54.

[23] S. Kuno and A. G. Oettinger, Multiple Path Syntactic Analyzer, Report No. NSF-8 to the National Science Foundation, Section 1, Harvard Computation Laboratory, Cambridge, January 1963.

[24] N. Chomsky, Aspects of the Theory of Syntax, M.I.T. Press, Cambridge, 1965.

[25] Y. Wilks, Transformational Grammars again, System Development Corporation Report SP 2936, August 29, 1967.

[26] Veillon, G., Veyrunes, J., and Vauquois, B., A Metalanguage of Transformational Grammars; Application to Problems of Transfer and Syntactic Generation, In 2ème Conférence Internationale sur le Traitement Automatique des Langues, Paper No. 22, Grenoble, August 1967.

[27] D. E. Walker, On-Line Text Processing — Introduction and Overview, MTP-57, Information System Language Studies, No. 12, Mitre Corporation, Bedford, March 1967.

References
(contd)

[28] R. F. Simmons and J. F. Burger, A Semantic Analyzer for English Sentences, System Development Corporation Report SP 2987, Santa Monica, 1967.

[29] A. J. Garfinkel, Semantic Representation Theory, System Development Corporation, Report TM 3687/000/00, Santa Monica, September 29, 1967.

[30] M. R. Quillian and D. G. Bobrow, Growing a Semantic Memory from English Text, Fifth Annual Meeting of AMTCL, Atlantic City, New Jersey, April 1967.

[31] A. Jolkovsky and I. Mel'chuk, Essai d'une Theorie Semantique Applicable au Traitement du Langage, 2ème Conférence Internationale sur le Traitement Automatique des Langues, Paper No. 30, Grenoble, August 1967.

[32] Y. Wilks, Semantic Consistency in Text — An Experiment and Some Suggestions, System Development Corporation Report SP-2758, Santa Monica, February 20, 1967, 38 pages.

[33] P. Stone, Improved Quality of Content Analysis Categories, Computerized Disambiguation Rules for High Frequency English Words, National Conference on Content Analysis, Philadelphia, November, 1967.

[34] P. A. W. Lewis, P. B. Baxendale and J. L. Bennett, Statistical Discrimination of the Synonymy/Antonymy Relationship Between Words, Journal of the ACM, Vol. 14, No. 1, January 1967.

[35] K. Sparck Jones, Notes on Semantic Discourse Structure, System Development Corporation Report SP-2714, Santa Monica, March 3, 1967.

[36] J. C. Olney, D. L. Londe, An Analysis of English Discourse Structure with Particular Attention to Anaphoric Relations, System Development Corporation Report SD-2769, November 10, 1967.

[37] W. A. Woods, Semantics for a Question-Answering System, Harvard University Doctoral Thesis, Mathematical Linguistics and Automatic Translation, Report NSF-19, Cambridge, September 1967.

[38] D. G. Bobrow, Problems in Natural Language Communications with Computers, IEEE, Transactions on Human Factors in Electronics, Vol. 8, March 1967.

[39] J. B. Fraser, On the Role of Natural Language in Man-Machine Communication, in Information System Science and Technology, W. D. Walker, editor, Thompson Book Co., Washington, 1967, pp. 21-2

[40] T. Givón, Notes on the Grammatical Characterization of English Queries, Technical Memorandum 3565/000/00, System Development Corporation, July 20, 1967.

References
(contd)

[41] S. Y. Kuroda, A Sketch of a Target Query Language, Mathematical Linguistics and Automatic Translation, Report NSF-18 to the National Science Foundation, Section 10, August 1967.

[42] L. Horn, Some Transformations for Answering English Questions, System Development Corporation, Report TM 3691/000/00, Santa Monica, October 4, 1967.

[43] Jane J. Robinson, The Transformation of Sentences for Information Retrieval, The Rand Corporation, Santa Monica, December 1965.

[44] P. S. Rosenbaum, A Grammar Base Question-Answering Procedure, Communications of the ACM, Vol. 10, No. 10, October 1967.

[45] F. M. Tonge, A Simple Scheme for Formalizing Data Retrieval Requests, Rand Memorandum RM-5150-PR, May 1967.

[46] E. B. Parker, SARPSIS: Syntax Analyzer, Recognizer, Parser and Semantic Interpretation System, Stanford Physics Information Retrieval System, September 29, 1967.

[47] H. Cautin and F. Rapp, Description of Easy English, Moore School of Electrical Engineering, Report 67-22, April 1967.

[48] Soergel, Dagobert, Mathematical Analysis of Documentation Systems — An Attempt to a Theory of Classification and Search Request Formulation, Information Storage and Retrieval, Vol. 3 No. 3, July 1967, pp. 129-173.

[49] M. Coyaud, N. Siot Decauville, L'Analyse Automatique des Documents, Mouton, Paris, 1967.

[50] G. Salton, Automatic Content Analysis in Information Retrieval, National Conference on Content Analysis, University of Pennsylvania, Philadelphia, November 1967.

[51] National Academy of Sciences, Language and Machines: Computers in Translation, (ALPAC Report) National Academy of Sciences, Washington, 1966.

[52] L. Peter Deutsch and B. W. Lampson, An Online Editor, Communications of the ACM, Vol. 10, No. 12, December 1967.

[53] G. Peitz and W. L. White, A console system for text manipulation and page composition, Progress Report No. 13, Bunker Ramo Corporation, March 20, 1967.

References
(contd)

[54] W. R. Sutherland, Language Structure and Graphical Man-Machine Communication, Information Systems Science and Technology, D. E. Walker, editor, Thompson Book Co., Washington, 1967, pp. 29-31.

[55] S. Kuno, Research and Development of a System for Inputting and Outputting Chinese Characters, unpublished notes, November 17, 1967.

[56] Klein, Sheldon, B. G. Davis, W. Fabens, R. G. Herriot, B. J. Katke, M. A. Kuppin, A. E. Towster, Autoling: An Automated Linguistic Fieldworker, in 2ème Conférence Internationale sur le Traitement Automatique des Langues, Grenoble, Paper No. 31, August 1967.

[57] J. K. Summers and E. Bennett, AESOP — A Final Report: a prototype on-line interactive information control system, in Information Systems Science and Technology. D. E. Walker, editor, Thompson Book Co., Washington 1967.

[58] W. V. English, D. C. Englehart and M. L. Berman, Display Selection Techniques for Text Manipulation, IEEE, Transactions on Human Factors in Electronics, Vol. 8, March 1967.

[59] W. Teitelman, Pilot, A Step Toward Man-Computer Symbiosis, M.I.T., Doctoral Thesis, Report MAC-TR-32, September 1966, 193 p.

[60] J. Weizenbaum, Contextual Understanding by Computers, Communications of the ACM, Vol. 10, No. 8, August 1967.

[61] Galli, E. J. and Yamada, H., An Automatic Dictionary and the Verification of Machine-Readable Text, IBM Systems Journal, Vol. 6, No. 3, 1967.

[62] C. H. Davis, An Approach to Automated Vocabulary Control in Indexes of Organic Compounds, Journal of Chemical Documentation, Vol. 7, August 1967, pp. 131-134.

[63] Olney, J., Revard, C. and Ziff, P., Toward the Development of Some Computational Aids for Obtaining a Formal Semantic Description of English, in 2ème Conférence Internationale sur le Traitement Automatique des Langues, Paper No. 33, Grenoble, August 1967.

[64] J. A. Bachrach and L. Hirschberg, Une Troisième Version du Dicautom, 2ème Conference Internationale sur le Traitement Automatique des Langues, Grenoble, August 1967.

[65] Masterman, Margaret, Man-aided Computer Translation from English into French using an On-Line System to Manipulate a Bi-lingual Conceptual Dictionary of Thesaurus, 2ème Conférence Internationale sur la Traitement Automatique des Langues, Paper No. 12, Grenoble, August 1967.

References
(contd)

[66] G. Salton, Information Dissemination and Automatic Information Systems, Proceedings of the IEEE, Vol. 54, No. 12, December 1966, pp. 1663-1678.

[67] P. J. Horvath, A. Y. Chamis, and R. F. Carroll and J. Dlugos, The B. F. Goodrich Information Retrieval System and Automatic Information Distribution using Computer Compiled Thesaurus and Dual Dictionary, Journal of Chemical Documentation, Vol. 7, August 1967, pp. 124-130.

[68] G. London, A classed thesaurus as an aid to indexing, Rutgers University, Graduate School of Library Science, October 1966.

[69]. M. Rubinoff, W. Franks and D. C. Stone, Description of an Experiment Investigating Term Relationships as Interpreted by Humans, Moore School of Electrical Engineering, June 1967.

[70] S. Y. Sedelow, Stylistic Analysis, Report on the Third Year of Research, System Development Corporation, Report TM-1908/300/00, March 1967.

[71] L. N. Gross, On-Line Programming System User's Manual, Information System Language Studies No. 14, The Mitre Corporation, March 1967.

[72] H. Hayashi, Snobol Ling Operation Manual, Aiken Computation Laboratory, Harvard University manuscript, September 6, 1967.

[73] M. Kay, Experiments with a Powerful Parse, 2ème Conférence Internationale sur le Traitement Automatique des Langues, Grenoble, 1967, also Rand Memorandum RM-5452-PR, October 1967.

[74] D. L. Londe and W. J. Schoene, TGT: Transformational Grammar Tester, Report TM 3759/000/00, System Development Corporation, November 17, 1967.

[75] J. F. Burger, R. E. Long and R. F. Simmons, An Interactive System for Computing Dependencies, Phrase Structure and Kernels, System Development Corporation, SD 2454/000/01, June 29, 1967.

[76] P. G. Chapin, L. N. Gross, L. M. Norton, R. J. Beller, C. T. Browne, SAFARI, an On-Line Text-Processing System User's Manual, Information Systems Language Studies No. 15, The Mitre Corporation, March, 1967.

[77] D. E. Walker, SAFARI - An On-Line Text Processing System, Proceedings of the 30th Annual ADI Meeting, Thompson Book Co., Washington, 1967, pp. 144-147.

References
(contd)

[78] D. J. Hillman, Document Retrieval Theory, Relevance, and the Methodology of Evaluation, Lehigh University, Reports No. 2, 3, 5, No. 2 Structure of Lecom, July 1966
No. 3 Microcategorization for Text Processing, July 1966.
No. 5 Arithmetization of Syntactic Analysis, July 1967.

[79] D. J. Hillman, Negotiation of Inquiries in an On-Line Retrieval System, First Cranfield International Conference on Mechanized Information Storage and Retrieval Systems, Cranfield, August 1967.

[80] A. J. Kasarda, A Syntactically Oriented Natural Language Document Retrieval System with a Browsability Feature, in "Experimental Retrieval Systems Studies", Lehigh University, Report No. 3, April, 1967.

[81] G. Salton, The Identification of Document Content: A Problem in Automatic Information Retrieval, Proceedings of a Harvard Symposium on Digital Computers and their Applications, Annals of Computation Laboratory, Vol. 21, Harvard University Press, 1962.

[82] P. B. Baxendale and D. C. Clarke, Documentation for an Economical Program for the Limited Parsing of English: Lexicon, Grammar and Flowcharts, IBM Research, Report RJ 386, August 16, 1966.

[83] S. Klein and R. F. Simmons, A Computational Approach to Grammatical Coding of English Words, Journal of the ACM, Vol. 10, No. 3, July, 1963.

[84] G. Salton and M. E. Lesk, Computer Evaluation of Indexing and Text Processing, Section III, ISR-12 to the National Science Foundation, Cornell University, June, 1967, also in Journal of the ACM, Vol. 14, No. 1, January 1968.

[85] Jessica S. Melton, Automatic Processing of Metallurgical Abstracts for the Purpose of Information Retrieval, Final Report NSF-4, Center for Documentation and Communication Research, Case Western Reserve University, Cleveland, Ohio, July, 1967.

[86] J. J. Rocchio and G. Salton, Information Search Optimization and Iterative Retrieval Techniques, Proceedings of the 1965 Fall Joint Computer Conference, Las Vegas, November 1965.

[87] A. Borillo, J. Virbel, Problèmes Syntaxiques de L'indexation Automatique de Documents, 2ème Conférence Internationale sur le Traitement Automatique des Langues, Grenoble, August 1967.

References
(contd)

[88] A. Borillo, M. Coyaud and J. Virbel, Etudes sur l'Indexation Automatique, Report 13 DGRST, Contract 65-FR-160, CNRS-SAD, June 1966.

[89] A. Borillo, N. Siot Decauville and J. Virbel, Etudes sur l'Indexation Automatique, Report 18 DGRST, Contract 65-FR-160, CNRS-SAD, December 1966.

[90] J. Virbel, Etudes sur l'Indéxation Automatique, Report 20 DGRST, Section d'Automatique Documentaire, Centre National de la Recherche Scientifique, June 1967.

[91] N. Sager, Report on the String Analysis Programs, Introductory Volume, Department of Linguistics, University of Pennsylvania, March 1966.

[92] M. Salkoff. Preliminary Report on the Application of the String Program to Information Retrieval of Physics Texts, Institute for Computer Research in the Humanities, New York University, January 1967.

[93] M. Salkoff, and N. Sager, The Elimination of Grammatical Restrictions in a String Grammar of English, 2eme Conference Internationale sur le Traitement Automatique des Langues, Paper No. 17, Grenoble, August 1967.

[94] P. E. Jones, V. E. Giuliano, and R. M. Curtice, Papers on Automatic Language Processing — Development of String Indexing Techniques Vol. 3, Report ESD-7R-67-202, A. D. Little, Inc., Cambridge, February 1967.

[95] C. Borkowski, An Experimental System for Automatic Recognition of Personal Titles and Personal Names in Newspaper Text, American Documentation, Vol. 18, July 1967, pp. 131-138, also 2eme Conference Internationale sur le Traitement des Langues, Grenoble, August 1967.

[96] R. Kittredge, Investigation of a Statistical Method for Segmentation of Sentences, Transformations and Discourse Analysis, Project No. 61, University of Pennsylvania.

[97] H. L. Resnikoff and J. L. Dolby, The Nature of Affixing in Written English - Part 2 MT, Vol. 9, June 1966, pp. 23-33.

[98] S. Y. Sedelow and W. A. Sedelow, Jr., Stylistic Analysis, in Automated Language Processing, Borko, editor, John Wiley and Sons, 1967.

References
(contd)

[99] M. Rubinoff and D. C. Stone, Semantic Tools in Information Retrieval, Proceedings of the 30th Meeting of the American Documentation Institute, New York, 1967.

[100] M. A. Libbey, Use of 2nd Order Descriptors for Document Retrieval, American Documentation, Vol. 18, No. 1, January, 1967.

[101] R. M. Curtice and D. E. Jones, Distributional Constraints and the Automatic Selection of an Indexing Vocabulary, Proceedings of the 30th Annual ADI Meeting, Thompson Book Co., Washington, 1967, pp. 152-156.

[102] S. F. Dennis, The Design and Testing of a Fully-Automatic Indexing - Searching System for Documents Consisting of Expository Text, in Information Retrieval - A Critical View, G. Schecter, editor, Thompson Book Co., 1967.

[103] B. E. Jones and R. M. Curtice, A Framework for Comparing Term Association Measures, American Documentation, Vol. 18, July 1967, pp. 153-161.

[104] V. E. Giuliano and P. E. Jones, Linear Associative Information Retrieval, in Vistas in Information Handling, P. Howerton, editor, Spartan Books, 1963.

[105] P. E. Jones, V. E. Giuliano, and R. M. Curtice, Papers on Automatic Language Processing — Linear Models for Associative Retrieval Vol. 2, A. D. Little, Inc., Report ESD-TR-67-202, Cambridge, February 1967.

[106] M. E. Lesk, Word-Word Associations in Document Retrieval Systems, Section IX, Report No. ISR-13 to the National Science Foundation, Department of Computer Science, Cornell University, January 1968.

[107] R. R. Anderson, An Associativity Technique for Automatically Optimizing Retrieval Results, in Experimental Retrieval System Studies, Lehigh University, Report No. 3, April 1967.

[108] C. G. Zull and G. J. Baumanis, Texture Correlates of Relevance Preliminary Investigation, CDCR, School of Library Science, WRU, CSL Techical Report No. 9, June 1967.

[109] K. Sparck Jones and D. Jackson, Current Approahces to Classification and Clump Finding at the Cambridge Language Research Unit, Computer Journal, Vol. 10, No. 1, May 1967.

[110] K. Sparck Jones and D. M. Jackson, The Use of the Theory of Clumps at the Cambridge Language Research Unit, Report ML 200, C.L.R.U., June 1967.

References
(contd)

[111] G. N. Lance and W. T. Williams, Notes on the Classification of Multi-Level Data, Computer Journal, Vol. 9, February 1967, pp. 381-382.

[112] L. B. Doyle, Breaking the Cost Barrier in Auto Classification, System Development Corporation, Report SP-2516, July 1, 1966.

[113] R. M. Needham, The Termination of Certain Iterative Processes, Rand Memorandum, RM 5188-PR, November 1966.

[114] B. Nelson, Machine Translation: Committee Skeptical over Research Support, Science, Vol. 155, No. 3758, January 6, 1967.

[115] O. S. Kulagina and I. A. Mel'chuk, Automatic Translation: Some Theoretical Aspects and the Design of a Translation System, from Machine Translation, A. D. Booth, editor, North Holland Publishing Co., Amsterdam, 1967.

[116] R. J. Solomonoff, Some Recent Work in Artificial Intelligence, Proceedings of the IEEE, Vol. 54, December 1966, pp. 1687-1697.

[117] P. L. Garvin, Heuristic Syntax - Computer Aided Research in Machine Translation, Bunker Ramo Corporation, Progress Report No. 14, March 27, 1967.

[118] D. B. Orr and V. H. Small, Comprehensibility of Machine-aided Translations of Russian Scientific Documents, MT, Vol. 10, No. 1-2, March, June 1967.

[119] J. J. Magnino, Jr., IBM's Unique but Operational International Industrial Textual Documentation System, 33rd Conference of FID, Tokyo, September 1967.

[120] P. J. Nelson, User Profiling for Normal Text Retrieval, Proceedings of the 30th Annual ADI Meeting, Thompson Book Co., Washington, 1967.

[121] J. H. Schneider, A Computerized Operational Selective SDI System based on a Hierarchical Decimal Classification of Biomedical Information, Proceedings of 30th Annual Meeting of the American Documentation Institute, October 1967.

[122] H. Borko, Utilization of On-Line Interactive Displays, in Information Systems Science and Technology, D. Walker, editor, Thompson Book Co., Washington, 1967, pp. 327-334.

[123] H. P. Burnaugh, The Bold Bibliographic On-Line Display System, in G. Schecter, editor, Information Retrieval — A Critical View, Thompson Book Co., 1967.

[124] W. D. Mathews, The TIP Retrieval System at M.I.T., in Information Retrieval — A Critical View, G. Schecter, editor, Thompson Book Co., 1967.

References
(contd)

[125] H. Cautin, T. C. Lowe, F. Rapp, and M. Rubinoff, An Experimental On-Line Information Retrieval System, University of Pennsylvania, Moore School of Electrical Engineering, April-May, 1967.

[126] T. C. Lowe, Design Principles for an On-Line Information Retreival System, Moore School of Electrical Engineering, Report 67-14, December 1966.

[127] M. Rubinoff, S. Bergman, W. Franks and E. R. Rubinoff, Experimental Evaluation of Information Retrieval Through a Teletypewriter, Moore School of Electrical Engineering, June 1967.

[128] M. Rubinoff, The Moore School Information Systems Laboratory, Moore School of Electrical Engineering, May 1967.

[129] G. Salton, Search Strategy and the Optimization of Retrieval Effectiveness, Section V, Scientific Report ISR-12 to the National Science Foundation, Department of Computer Science, Cornell University, June 1967.

[130] G. Salton, The SMART Project — Status Report and Plans, Section 1, Scientific Report ISR-12 to the National Science Foundation, Department of Computer Science, Cornell University, June 1967.

[131] J. A. Craig, S. C. Berezner, H. C. Carney, C. R. Longyear, Deacon: Direct English Access and Control, Proceedings of the AFIPS Fall Joint Computer Conference, 1966, Thompson Book Company, Washington, 1966, pp. 365-380.

[132] Stephen L. Coles, Syntax Directed Interpretation of Natural Language, Doctoral Thesis, Carnegie Tech, Pittsburgh, Penna., 1967.

[133] C. H. Kellogg, Converse — A system for the on-line description and retrieval of structured data using natural language, System Development Corporation, SP-2635, Santa Monica, May 1967.

[134] C. H. Kellogg, On-Line Translation of Natural Language Questions into Artificial Language Queries, System Development Corporation, SP-2827, Santa Monica, April 28, 1967.

[135] R. E. Levien and M. E. Maron, A Computer System for Inference Execution and Data Retrieval, Communications of the ACM, Vol. 10, No. 11, November 1967.

[136] A. Kasher, Data Retrieval by Computer — A Critical Survey, in The Growth of Knowlsdge, M. Kochen, editor, John Wiley and Sons, New York, 1967.

References
(contd)

[137] C. Cleverdon, J. Mills and M. Keen, Factors Determining the Performance of Indexing Systems, Vol. 1: Design, Vol. 2: Test Results, Aslib Cranfield Research Project, Cranfield, 1966.

[138] F. Alouche, N. Bely, R. C. Cros, J. C. Gardin, F. Levy, J. Perriault, Economie Generale d'une Chaine Documentation Mecanisee, Gauthiers Villan, Paris, 1967.

[139] T. Saracevic and L. Rothenberg, Procedure Manuals for the Comparative Systems Laboratory, CDCR, School of Library Science, WRU, March 1967.

[140] T. Saracevic and J. Hazelton, Operational Procedures for Question Analysis, in Procedure Manuals for Comparative Systems Laboratory, WRU-CDCR, March 1967.

[141] T. Saracevic, Operational Procedures for Search Strategy in Procedure Manuals for the Comparative Systems Laboratory, Center for Documentation and Communications Research, Comparative Systems Laboratory, Cleveland, March 1962.

[142] T. Saracevic and A. H. Rees, Towards the Identification and Control of Variables in Information Retrieval Experimentation, Journal of Documentation, Vol. 23, March 1967, pp. 7-19.