> "Te Nuyl...uses, as quasi-descriptors, word-sets chosen from the Oxford English Dictionary (e.g., any word falling between A-Ah) and relies on the subsequent correlation of terms to make sense of his seemingly bizarre choice." 1/

Lefkovitz is concerned with the so-called "automatic stratification" of a file in which both generic or associative relationships and exclusive partitioning is used to facilitate search. He claims:

> "...The exclusive partitioning implies a separation of descriptors into groups such that no two descriptors in a group co-occur in any given document description of the file. This arrangement presents the dissociative properties of the file, or forbidden combinations. When coupled with a superimposed display of the 'inclusive' or associative properties of the file a unique classification of the descriptors of this file results, which is based solely upon the association of the descriptors themselves within the document descriptions and not upon an arbitrary set of classes constructed by professional indexers." 2/

The purpose is to assist the searcher by warning him that if he chooses more than one descriptor from any one group as terms in his search request, there will be a null response from this particular file. However, the particular application considered involves a limited number of highly quantifiable or scalable "attribute-value" pairs, (for so the descriptors involved are defined), such as "Age-23", and "Hair-red". It is by no means obvious that comparable exclusive partitionings could be achieved for literature items or that the recomputations necessary as new items enter the file can be achieved on a practical basis.

6. OTHER POTENTIALLY RELATED RESEARCH

In this section we shall consider certain areas of potentially related research that may prove applicable to the improvement of automatic indexing techniques. First is the area of thesaurus construction and use, which in turn is somewhat related to the development of statistical association techniques, especially for "indexing-at-time-of-search" and search renegotiations. Natural language text searching will also be briefly considered, together with related research in the general area of linguistic data processing.

6.1 Thesaurus Construction, Use, and Up-Dating

The first area of potentially related research which promises improvements in automatic indexing procedures is that of thesaurus lookups by machine. There are several different possible definitions of the word "thesaurus" in the context of information storage, selection and retrieval systems. The first is that it is a prescriptive indexing aid, or authority list, serving the function of normalizing the indexing language, primarily by the use of a single word form for words occurring in various inflections, by the reduction of synonyms, and by the introduction of appropriate syndetic devices. The second definition relates to the intended function for the provocation and suggestion to the indexer or the searcher of additional terms and clues, and it follows the idea of word groupings related to concepts as in a traditional thesaurus like Roget's. The third

1/ Cleverdon and Mills, 1963 [131], p. 8.

2/ Lefkovitz, 1963 [353], Preface, pp. VIII-IX.

possible definition involves the special case of devices or techniques which display or use prior associations and co-occurrences or words, indexing terms, and related documents to provide a guide or suggestive indexing and search-prescription-formulation or renegotiation aid.

The idea of a mechanized authority list, following the restrictive first definition, has been proposed by a number of investigators [1] and has actually been used in computer programs as discussed for example by Schultz and Shepherd (1960 [532]), Shepherd (1963 [545]) and Artandi (1963 [20]). It is the second definition of thesaurus with which we shall be principally concerned. It is, as we have said, close to the conventional idea of such a thesaurus as Roget's. It is based on the hypothesis that patterns of co-occurrences of words in a new item or in a search request can be compared with patterns of prior co-occurrences, as given by a thesaurus "head", in order to expand, clarify, or pin-point "meaning" and thus provide a more effective indication of the true subject content. The third definition will be considered as falling within the more general scope of statistical association techniques, although as Giuliano points out, "a retrieval system embodying an automatic thesaurus thus qualifies as being 'associative'." [2]

The application of a thesaurus-like approach to indexing and searching problems is again an area in which Luhn is one of the earliest proponents. In January 1953, he proposed a new method of recording and searching information in which a special dictionary would be compiled for use in broadening the terms of a search request and in normalizing word usage as between various indexers (recorders) and searchers. Although he did not then use the term "Thesaurus" as such, he said in part:

> "The process of broadening the concept involves the compilation of a dictionary wherein key terms of desired broadness may be found to replace unduly specific terms, the latter being treated as synonyms of a higher order than ordinarily

[1] See, for example, "Summary of discussions, Area 5," ICSI, 1959 [578], p. 1263: "Two further complications arise from a mechanical index. Some articles might deserve as an indexing term a word not contained in the article. By an authority list, the product of the mechanized indexing procedure might have such additional words added to it. Again, an article might use a particular word but the vocabulary of the system might prefer another one. This also can be handled by a mechanized authority list.".

[2] Giuliano and Jones, 1962 [229], p. 4.

> considered. Translating criteria into these key terms is a process of normalization which will eliminate many disagreements in the choice of specific terms amongst recorders, amongst inquirers, and amongst the two groups, by merging the terms at issue into a single key term. However, the dictionary does not classify or index but maintains the idea of being fields... A specific term may appear under the heading of several key terms and if according to its application an overlapping of concepts exists then the term is represented by the several key terms involved..." [1]

In subsequent papers, Luhn has developed related ideas of a "family of notions" and "dictionaries of notional families". [2] In particular, he emphasizes that for automatic indexing, by contrast with automatic abstracting, consideration should be given to the normalization of variations in author-chosen terminology: "It will be necessary for a machine to resolve variation of word usage with the aid of a device the functions of which resemble a dictionary at one level and of a thesaurus at another level of requirements." [3]

The first issue of the National Science Foundation's compendium of project statements, "Current Research and Development in Scientific Documentation", which appeared in July 1957 [430] reported several projects of interest in terms of thesaurus construction and use, [4] namely: (1) work by Luhn at IBM involving the establishment of a thesaurus to facilitate encoding of items whose texts would be available in machine-usable form, (2) work by Bernier and Heumann at Chemical Abstracts Service looking toward the development of a technical thesaurus, (1957 [57]), and (3) an approach to mechanized translation proposing to use a mechanized thesaurus at the Cambridge Language Research Unit. This latter project incorporated the ideas of Masterman and her associates from about 1956 on (Halliday 1956 [249], Masterman, 1956 [403]; Joyce and Needham, 1958 [305]), to apply the principle of checking co-occurrences of text words against thesaurus "heads" to which they belonged, in order to resolve homographic ambiguities and thus achieve more idiomatic translation by machine.

For the ICSI Conference in 1958, Masterman, Needham and Sparck-Jones prepared a paper discussing analogies between machine translation and information retrieval, and recapitulated the arguments of Needham and Joyce for the use of a thesaurus in the formulation of search requests, as follows:

> "If a large number of terms are used to describe a document, the existence of synonyms is likely: in a system such as uniterm no attempt is made to bracket the synonyms, which means that a request will produce only the document described

[1] Luhn, 1953 [383], p. 15.

[2] Luhn, 1959 [371], p. 51, 1959 [384]; 1957 [385], p. 316.

[3] Luhn, 1959 [384], p. 12.

[4] National Science Foundation's CR&D Report No. 1, [430], pp. 21, 6, 4.

> in identical terms and not in synonymous ones. If the existence of synonyms is avoided, by using a small number of exclusive descriptors, the description of a document in terms useful for retrieval is more difficult, also it is equally difficult to relate a request to the description of documents. A further difficulty is that descriptions only list the main terms, and take no account of their relations to one another. The C. L. R. U. experiments being carried out make use of a thesaurus, a procedure through which it is hoped that these difficulties will be avoided and that a request for a document although not using the same terms as those in the document will produce that document and others dealing with the same problem, but described in different, though synonymous, terms." [1]

In general, the use of a thesaurus to constrain variations in word or term usage (as in our first definition, a mechanized authority list), to reduce synonymity, to resolve homographic ambiguity, to provoke and suggest additional terms or ideas to indexer and to searcher alike, is related to the improvement of automatic indexing procedures in precisely the same sense that its use would be effective in any indexing system whatsoever. In another sense, however, the construction and use of the thesaurus is related to linguistic data processing by machine in another way. Garvin suggests:

> "...One may reasonably expect to arrive at a semantic classification of the content-bearing elements of a language which is inductively inferred from the study of text, rather than superimposed from some viewpoint external to the structure of the language. Such a classification can be expected to yield more reliable answers to the problems of synonymy and content representation than the existing thesauri and synonym lists, which are based mainly on intuitively perceived similarities without adequate empirical controls." [2]

This is with respect to the recognition that the machine itself can be used to compile and construct the thesaurus. While Luhn in some of his 1957-8 proposals still considered the compilation and organization of a thesaurus to be primarily a matter of human effort, he nevertheless pointed out that: "The statistical material that may be required in the manual compilation of dictionaries and thesauri may be derived from the original texts in any desired form and degree of detail." [3] De Grolier makes the complementary statement that the Luhn techniques should "considerably facilitate" the preparation of thesauri. [4]

Even more importantly, the computer can be used for periodic up-datings and revisions. The work on the FASEB index-term normalization procedures involved early recognition of the need to "educate the thesaurus" by examining print-outs when no matches occurred and providing a continuous process of amendment. [5] Computer-maintained statistics of word and term usages are closely related to possibilities for

[1] Masterman, Needham, and Sparck-Jones, 1958 [405], p.934-935; Needham and Joyce 1958 [305].

[2] Garvin, 1961 [224], p. 138.

[3] Luhn, 1959 [354], p. 12.

[4] De Grolier, 1962 [152], p. 132.

[5] Shepherd, 1963 [545], p. 392.

construction and revision of a mechanized thesaurus, as again Luhn has suggested. [1] Schultz suggests that machine records should be maintained of what thesaurus terms are actually used for indexing and searching, the frequencies of term usage, the co-occurrences, the number of items described by particular combinations of terms and the like.[2]

The potential combinations of natural text processing, automatic indexing, and thesaurus construction and updating are stressed in many current programs. For example, Eldridge and Dennis discuss:

> "Indexing by machine from natural text in a fully automatic system, in which statistical analysis of the words is employed as a device for (a) building automatically a 'concept' thesaurus, (b) indexing incoming documents with reference to the thesaurus, and (c) continuously revising the thesaurus to reflect new word usages in currently incoming documents."

Similarly, Giuliano and Jones suggest that given a term-term statistical association matrix, a transformation can be arrived at with a unit vector assigning value only to index term Z that ranks every other index term according to degree of association with Z, then by listing the higher ranked terms for each term Z, "a 'thesaurus' listing can be obtained completely automatically." [4]

6.2 Statistical Association Techniques

A special definition of the word "thesaurus" might, as we have noted, include the development of devices and techniques which either automatically or by man-machine interaction serve to suggest the amplification of a set of index terms. We shall briefly consider here both devices that visually display associations between words, terms, and documents [5] and techniques for machine use of coefficients of correlation for prior co-occurrences in a collection of word-word, word-term, term-term, term-document, and document-document associations, the statistical association factor technique as first developed by Stiles.

[1] Luhn, 1957 [385], p. 316: "Provision should be made to register the number of times each word is looked up in the index and the number of times each family number has been used for encoding. Such a record would be an indispensable part of the system for making periodic adjustments based on the usage of words or notions as mechanically established."

[2] Schultz, 1962 [529], p. 104.

[3] Eldridge and Dennis, 1962 [183], p. 6.

[4] Giuliano and Jones, 1962 [229], p. 12.

[5] It should be noted that Tabledex, the Scan-Column Index, and similar tools provide to some extent a display of prior associations between index terms. (See pp. 25-27 of this report.) Thus Cheydleur (1963 [115], p. 58) remarks: "Ledley. . has focussed on inter-item concepts in designing his economical TABLEDEX arrangement for displaying the connectivity of index terms and related file items."

6.2.1 Devices to Display Associations: EDIAC

The interest aroused among some documentalists by the provocative idea of a "Memex" to record and display associations between ideas as proposed by Bush in 1945 ([93]) led to specific attempts at Documentation, Inc. in the 1950's to develop a device which would incorporate at least the associations between indexing terms assigned to documents and between documents with respect to their sharing of common indexing terms (1954 [157], 1956 [155, 156]). The first approach to this objective, as reported by Taube, was the idea of a manual dictionary of terms arranged in alphabetical order, with a "page" reserved for each and every indexing term used for any document in the collection. On each page would be listed all other terms that had co-occurred with that term in the indexing of one or more documents. Another idea was to display associations of terms used in a collection through the "superimposition of dedicated positions in a set of cards or plates..." [1]

Subsequently, an actual device to demonstrate a system for display of term-term, term-document, and document-document associations, was built under an Office of Naval Research contract. [2] The demonstration model contained a vocabulary of 250 terms which had been used in various combinations to index 100 reports. Interconnections in an electrical network provided the associational linkages. A display panel was provided with symbol-indicators which could be lighted up to identify particular terms and particular report numbers.

This EDIAC device (for Electronic Display of Indexing Association and Content) was intended for use both in guiding an indexer to either the extension or refinement of his initial choice of indexing terms and in assisting the searcher. It was claimed that the operation of such a device would be extremely simple. Thus:

> "For the index question the searcher selects any term in which he is interested and applies a voltage. He is told instantly the number of the reports dealing with that subject. Putting voltage in at any term also lights all other terms associated with the first term..." [3]

A later analog device, ACORN, will be discussed below in connection with the work of Giuliano and associates, at Arthur D. Little, Inc.

6.2.2 Statistical Association Factors - Stiles

The name of H. Edmund Stiles, like those of Luhn, Baxendale, Maron, Swanson, Edmundson and Wyllys, is generally associated with pioneering innovations in those areas of mechanized documentation which are directly related to the use of high-speed computer capabilities. While Stiles' work has been directed primarily to problems of search prescription formulation and renegotiation based on the results of preliminary search, he has specifically recognized that the use of statistical word association techniques in searching operations can provide a logical corollary to automatic indexing procedures. Thus:

[1] Taube et al, 1954 [599], p. 102.

[2] It is described and illustrated in Taube et al, 1956 [599], p. 63 ff.

[3] Documentation, Inc. 1956 [156], p. 7.

> "Automatic indexing, based on the relative frequency of words used in a document, produces a partial vocabulary of the content words used to express its subject. Retrieval can then be accomplished by expanding the request vocabulary... This method tends to overcome the deficiencies and inconsistencies inherent in the use of terms derived automatically from a text." [1]

Conversely, Stiles also points out the possibility that the results of automatic derivative indexing procedures, extracting indexing words from the documents directly, might prove a more realistic or reliable basis for the development of his word co-occurrence correlation data than do the Uniterms assigned by human indexers. [2] The work of Stiles has also stressed the importance of two factors that may well be critical for the improvement of automatic indexing techniques. These are, namely, the consensus of prior human indexing and the consensus of subject coverage of a particular collection. [3]

In his experimental investigations, Stiles began with an existing collection of approximately 100, 000 items which had previously been indexed, over a period of time, with a Uniterm indexing vocabulary consisting of about 15, 000 terms. The objective of the experiments was to determine how, given a specific search request, a more effective "net to catch documents" [4] could be generated and how the responding items might be ranked in order of their probable relevance to the request.

The statistics of co-occurrence of terms used to index the same documents were first obtained. A modified chi-square formula was then applied to determine relative frequencies of use of co-occurring terms. [5] Patterns of term co-occurrence could then be derived in the sense of term-profiles which show, for each term, the more significant of its associational values of pairing with other terms in the collection. The actual procedure for using these term-profiles in search prescription formulation and in document selection involves several steps, generally as follows: [6]

[1] Stiles, 1962 [573], pp. 12-13.

[2] Stiles, 1961 [572], p. 205.

[3] Stiles, 1962 [573], p. 6 and 1961 [572], pp. 273, 277.

[4] Stiles, 1961 [572], p. 192.

[5] In general, we shall not be concerned with the precise mathematical formulations. It is to be noted that in a recent report Giuliano and his colleagues have reviewed a number of the various mathematical formulas proposed in the literature for the computation of word, term, and document associations, including those of Parker-Rhodes and Needham, Maron and Kuhns, Stiles, Salton, Osgood, Bennett and Spiegel (Giuliano et al, 1963 [230], Appendix I).

[6] Stiles, 1961 [571], pp. 273-275.

1. For each term in the initial formulation of a search request, the appropriate term-profile is obtained, which gives weighted values for those other terms that had significantly co-occurred with it.

2. The profiles of each term in a multi-term request are compared and those additional terms common to all or a specified number of the profiles are selected and added to the initial set.[1]

3. The "first generation" terms resulting from step 2 are next treated as though they also were request terms, and steps 1 and 2 are repeated for them.

4. A selection is made from some reasonable proportion of the profiles associated with the first generation terms to produce the "second generation" terms.[2]

5. The expanded list of search terms is then compared with the index terms assigned to each document in the collection, and whenever a match is found the weight of the request term is assigned to the matching document term. These weights are then summed to provide a numeric measure of probable document relevance to the original request.

6. Documents responding to the expanded request are printed out in the order of document relevance scores.

Some experiments have been made using a computer program which accepts up to 300 weighted terms in an expanded request vocabulary. Representative results have been reported, in part, as follows:

> "...We asked a qualified engineer to examine these documents and specify which were related to 'Thin Films' and which were not... This engineer was not familiar with our project...yet...we found a remarkably high correlation between his evaluation and the document relevance numbers... We then checked to see how the documents containing information on 'Thin Film' had been indexed. We found that the first five documents on our list had been indexed by both 'Thin' and 'Film'. Three more documents had been indexed by 'Film' alone, and other related terms. Two documents had not been indexed by either 'Thin' or 'Film', but only by a group of related terms, yet they contained information on 'Thin Films' and had a high document relevance number. By using association factors and a series of statistical steps, easily programmed for a computer, we were thus able to locate

[1] These are called "first generation terms" and tend to reflect only statistical associations without including synonyms and near-synonyms which, over the course of time, have occurred in the indexing vocabulary.

[2] Stiles, 1961 [571], p.274: "Among these we find words closely related in meaning to the request terms." An example given in Ref. [572], pp. 200-201, is the derivation of 'weathering,' 'fungicidal', 'deterioration', and 'preservatives' as second generation terms when the initial request included the terms 'plastics', 'fungus', 'coating', and 'tests'.

documents relevant to a request even though the documents had not been indexed by the terms used in the request." [1]

In another case, which was analyzed in detail, a request profile of 26 terms that had been intuitively weighted by the customer resulted in the machine listing of 246 presumably responsive documents. Of these, 81 documents were of primary interest to the customer, and an additional 78 were of secondary interest to him. [2]

The statistical association technique as proposed by Stiles has also been investigated at the Datatrol Corporation, with particular reference to the field of legal literature (Hammond et al, 1962 [251]). About 350 documents in the field of Federal public law were indexed in cooperation with George Washington University, using a vocabulary of 680 index terms. A computer program was written for the IBM 7090 that can accommodate a 1200 x 1200 matrix to calculate the Stiles' association factors. Trials were made of various thresholds to determine which other terms were sufficiently high in association strength to a particular term to be selected for that term's profile.

Given the generation of the term profiles, a less sophisticated computer such as the 1401 can be used for the expansion of request terms and the actual conduct of searches. Such a program was demonstrated at the Annual Meeting of the American Bar Association, August 1962, with running of "live" requests suggested by jurists and with what are claimed to be "highly gratifying results". A point of interest relates to the question of updating of term-profiles and other statistical association factor data. Hammond, et al report:

> "The term profiles were generated a total of three times in the course of the pilot study, making it possible, to some extent, to assess the effect of vocabulary growth. Judging from this limited experience, it appears that a bi-monthly, or perhaps even quarterly, recompilation of term profiles should be sufficient for a mature collection." [3]

6.2.3 The Association Map - Doyle and Related Work at SDC

The name of Doyle is again that of an early and prolific investigator and innovator in the field of mechanized documentation and linguistic data processing. One of his provocative suggestions is generally known, in his own terminology, as that of "semantic road maps for literature searchers" or an "association map" technique. As a matter of convenience, we have chosen to consider this suggestion and a variety of related work

[1] Stiles, 1961 [577], pp. 198-199.

[2] Stiles, 1962 [573], p. 9.

[3] Hammond et al, 1962 [251], p. 6.

under the general heading of the association map technique, [1] although passing reference has been made to some of Doyle's suggestions and findings elsewhere in this report.

Beginning in 1958 (Doyle, 1959 [168]) information retrieval projects at the System Development Corporation have had, among other objectives, that of developing ways to use computers in the processing and interpretation of natural language text. By February of 1959, a computer program was already in operation that could search fragments of about 100 words of keypunched text, match input words against a pre-established clue word selection list (i. e., an inclusion dictionary) and substitute a short encoded form to be used for subsequent search. Processing of keypunched abstracts using this program involved computer time at the rate of four abstracts per second.

Other features of this text compiler, and of subsequent text processing programs developed at SDC, enable the making of frequency counts and other statistical measures. Such features are then used for the investigation of, for example, word-word, word-document, and word-subject associations, looking toward the determination of answers to such questions as: "Do subject words have distribution characteristics within a library that a computer program can detect?" [2]

Doyle's investigations of word co-occurrences have included hypotheses and tests of various probabilistic measures in terms of observed frequencies, in terms of "boing!" words (so-called because of the mental sound effect they elicit), [3] in terms of adjacent word pairs and affinities between particular nouns and particular adjectives, [4] and in terms of distinctions between frequency (the total number of times a word appears in a given library corpus) and prevalence (the total number of items in which a particular word appears). [5] He has also stressed distinctions between adjacent words and high correlations for words that are not closely positioned together in text, as follows:

[1] Compare Doyle himself, 1962, [163], p. 383: "Swanson and others have offered thesauri of synonyms and related terms... (to assist in indexing or search processes)... An association map is, in a sense, an extension of this solution; it is a gigantic, automatically derived thesaurus. Confronted by such a map, the searcher has a much better 'association network' than the one existing in his mind, because it corresponds to words actually found in the library, and, therefore, words which are best suited to retrieve information from that library." See also Wyllys, 1962 [651], p. 16: "L. B. Doyle (1961) has invented a fascinating search tool which seems to us to belong at a level intermediate between automatic indexes and automatic abstracts; i. e., a possible search method might be to have the computer scan automatic indexes and compare the index terms therein with the request, then obtain the possibly pertinent documents and display their association map for the user to examine..."

[2] Doyle, 1959 [168], p. 6.

[3] Doyle, 1959 [165], p. 5.

[4] Doyle, 1961 [169], p. 12; 1959 [165], p. 16.

[5] Doyle, 1962 [163], p. 380.

"We have also perceived that two different cognitive processes seem to be responsible for each type of correlation, one (adjacent correlation) involving the habitual use of word groups as semantic units, and the other (proximal correlation) having to do with the pattern of reference to various aspects of that which is being discussed. We can call the statistical effects, respectively, 'language redundancy', and 'reality redundancy'. Such a resolution of statistical effects is full of significance for information retrieval because it appears likely that reality redundancy can vary greatly from one science to another, whereas language redundancy, a universal property of talking and writing, is relatively invariant." [1]

With respect to the "semantic roadmap" or "association map" technique itself, Doyle's suggestion is that various measures of word and index term cross-associations may be applied to the generation of graphic displays of both types of co-occurrence relationships. Because of the variety of, in particular, the "proximal" correlations, it is assumed that the literature searcher should be given a display in which the representation of the assemblage of the varied relationships is two-dimensional rather than one. [2] An example is given, based upon computer processing of 600 abstracts of SDC internal reports to find intersections between 500 topical words, of associational connections for the word "output". This was generated by selecting the eight words most strongly correlated in the data with "output", such as "manual" and "radar", and then finding three other words highly correlated with each of these and also correlated with "output" itself. From the initial graph, it is further shown that item surrogates might be generated by word selection rules applied to documents to pick up, for example, "New York Air Defense → system → data → outputs → D.C. [3]

Continuing related work by Doyle and others at SDC has included various experimental studies of "pseudo-documents" consisting of lists of the twelve most frequently occurring words in 100-item samples of abstracts in various subject fields (Doyle, 1961 [161]). Of special interest in terms of potential improvements and modifications to machine indexing techniques are studies, based on similar lists, looking to the separation of words that may have been used in several different senses, i.e., the detection of homographs by statistical means (Doyle, 1963 [171]). More recent investigations by Doyle involve considerations of differences between word-grouping and document-grouping techniques and of possibilities for use of hybrid methods.

6.2.4 Work of Giuliano and Associates, the ACORN Devices

A program directed toward the design of "an English command and control language system" under an Air Force contract with Arthur D. Little, Inc., involves several interrelated aspects of natural language text processing, use of statistical association factors in search, man-machine interaction during search, and display of associational relationships by means of analog network devices. In this program and in related research, Giuliano and his associates are convinced that:

[1] Doyle, 1961 [169], p. 15.

[2] Doyle, 1962 [163], p. 379.

[3] Doyle, 1961 [169], pp. 24-25.

> "Automatic index term association techniques are needed to improve the recall of relevant information, to enable indexers and requestors to use language in a more natural manner, and to enable retrieval of relevant messages which are described by different index terms than those used in the inquiry." [1]

For the most part, the work to date has been directed to "associative retrieval" of messages limited to single sentences of English text, and to the search phases of a proposed system.

In the case of a corpus consisting of 230 sentences from a single text, a partially automatic indexing method was used. The text was first processed against a modified version of the Harvard Multipath Syntactic Analysis computer program and the resulting analyses were manually screened to select a unique, correct analysis for each sentence. Next, approximately 500 words, those that had been marked "noun" by the syntactic analyzer, were listed out and these in turn were manually screened to provide an "inclusion list" of 273 words. Sentences were then "indexed" with respect to which of these selected words they contained. Word associations were computed both in terms of co-occurrence within a sentence and of co-occurrence in syntactic structures.

Retrieval tests were then applied using both computer programs and the analog device, and evaluations were made on the basis of examining sentences selected in order of machine-ranked relevance and of comparisons of word lists associated with a given search term against association lists for another term picked at random. It is noted that, "although quantitative conclusions cannot be drawn", the results support the conclusion that: "Items retrieved due to automatically-generated associations tend to be more relevant than is explainable on a chance basis." [2]

The "request reformulation" retrieval program has also been used to generate term profiles from a collection of approximately 10,000 documents (previously indexed with at least 6 terms from a selective term vocabulary of 1,000 terms) which have then been compared against lists provided in the entries for corresponding terms in the Thesaurus of ASTIA Descriptors, Second Edition. The machine-produced association lists, at least for those words occurring relatively frequently in the corpus, appear to give thesaurus entries that are extensive, specific, and intuitively acceptable, and of high quality, especially with respect to listings of synonyms as well as factually related words. [3]

The development of the ACORN (Associative Content Retrieval Network) devices has provided additional tools for testing and display (1962 [229], 1963 [227, 304]). These devices are networks of passive resistance elements. Each word or index term and each sentence (240 by 230 in ACORN-IV) are represented by terminals interconnected by resistors with conductance equal to the connection strength, and with "leak" resistors

[1] Giuliano 1962 [228], p. 10.

[2] Giuliano et al, 1963 [230], p. 47.

[3] Ibid, pp. 57-58.

providing for various normalizations that may be applied to compensate for word or sentence frequency factors. These devices differ from the earlier EDIAC in the variable weightings provided, in the normalizations that may be applied, and in multipath interconnections.

When, for example, currents are applied at some of the word terminals, the voltages appearing on any of the other word terminals depend on the strengths of association between these words and the input words via all direct and indirect paths. The responses of sentence terminals to the input words of a query similarly depend upon how strongly a sentence is connected to these words and how strongly it is connected to other words which in turn are strongly connected to the query words. It is to be noted further that:

> "Pulling out or cutting a few randomly selected wires in an ACORN generally has a surprisingly small effect... This insensitivity is of course, explainable in terms of the multiplicity of indirect and redundant association paths which remain intact when a direct path is severed... It... suggests that the retrieval process can indeed be made insensitive to minor variations in indexing." [1]

In addition, there are intriguing possibilities for imposing a "viewpoint" with respect to a search by injecting bias currents. Thus if only non-"Air Force" jet plane items were desired, the "Air Force" items could in effect be grounded out. If there were no jet items in the collection other than those which were also Air Force items, these would be indicated as responsive, but largely they would appear only if this should be the case. Some words used have some connection to almost all other words, but these have little effect in the system and the hardware thus tends to compensate for the high frequencies of very general words.

6.2.5 Spiegel and Others at Mitre Corporation

Bennett and Spiegel, reporting at the Symposium on Optimum Routing in Large Networks, IFIP Congress-1962, [2] consider modifications to formulas for the calculation of statistical association factors which will normalize against such influences as frequency of word occurrences, relative word position within a string of words, and string length. This work has been carried forward at the Mitre Corporation in a program for developing procedures to encode various statistical properties of messages or documents and to use these codes for message routing and retrieval.

Differences between this approach and those of Maron and Kuhns, Stiles, and Doyle, relate primarily to the questions of how best to normalize. The objective is closely similar: to use associational weighting so as to provide, in response to a query, output of documents or messages ranked in order of probable relevance to the query.

[1] Giuliano and Jones, 1962 [229], p. 22.

[2] See Juncosa, 1962 [306], especially paper 4, E. Bennett and J. Spiegel, "Document and message routing through communication content analysis", pp. 718-719.

Additional features include provision for the matrix of coefficients of association to change with time or with deliberate manipulation to improve performance. Thus:

> "Each normalized cell weight... rises and falls with time as each specific association increases or decreases in relative frequency. In this way, the matrix memory of associations changes with time, maintaining a cumulative pattern of associations reflecting one statistical characteristic of messages fed into it in the past...
>
> "In addition to this adaptive characteristic of changing memory with time and with changing inputs, the matrix is also readily subject to formal education. Any specific cell weight can be strengthened by repeatedly reading into the matrix memory the specific strings that contain the desired associations. For example, by introducing the strings is am, is are, am is, am are, and are am, we can increase the statistical tendency of the tokens is, am, and are to be associated." [1]

Experimental results have been obtained for a corpus of 500 bibliographic entries contained in DDC's Title Announcement Bulletin. In the case of a three-term query, 40 items were selected and ranked in probable relevance order, with selection based on a particular relevance score value threshold. The investigators then reviewed the abstracts of all 500 items and rated them as to relevance with respect to the query. Seven additional items were found, of which three would have been machine-selected with a less stringent selection threshold. For the remaining four, it is reported that they "were poorly indexed and could have been judged not relevant by a human who depended upon the descriptor string only, as the matrix did, rather than upon review of the abstracts." [2]

6.3 Clues to Index-Term Selection from Automatic Syntactic Analysis

Several of the organizations and research teams most active in the investigation of linguistic data processing techniques, especially for automatic indexing, extracting and search renegotiation applications, are actively considering the use of clues derived from automatic syntactic analysis to improve criteria for machine selection of "significant" words, phrases, and sentences from raw text. Such approaches, in general, however, are subject to the limitations of non-availability of sufficient corpora of text in machine-usable form, in the first place, and, even more importantly by the non-availability of satisfactory computer programs for complete syntactic analysis up to the

[1] Spiegel et al, 1963 [566], p. 17.

[2] Ibid, p. 34.

present time. [1/] In terms of the state-of-the-art of automatic indexing, therefore, we shall not consider these approaches as more than indications for future research. A few suggestive examples are discussed briefly below.

The multi-pronged attack on mechanized information selection and retrieval problems headed by Salton and his associates includes the exploration of tree structures, to represent both the relationships between terms in a classification schedule or indexing term vocabulary and the representation of the results of automatic syntactic analyses of natural language text. It is proposed, then, that computer programs can achieve transformations of the syntactic trees representing word strings in the original text into simplified, condensed structures with normalized terms and can compare these trees with the classificatory trees (Salton, 1961 [516]). Manipulation of such trees together with appropriate dictionaries or thesauri can result, for a given proposed index term, in the finding of a preferred term for a particular system, or a set of synonymous terms, or sets of all terms in which the given term is included, and the like.

Anger considers some of the problems involved in complete syntactic analysis of texts with the objective of identifying the total network of relationships expressed and implied, as proposed by Lecerf, Ruvinschii, and Leroy, among others, of the Research Group on Automated Scientific Information (GRISA), EURATOM. Assuming that computer programs for syntactic analysis are or will be available, he suggests that simplifications may be obtained by determining only the basic relations that are indicated by direct syntactic dependencies or by linking words, (Anger, 1961 [15]).

A specific program for automatically extracting syntactic information from text has been studied by Lemmon (1962 [354]). The possibilities for combining dictionary lookups, word suffixes as indicators of syntactic role, and predictive syntactic analysis for text processing have also been further explored by Salton himself (1962 [518], 1963 [519]). A variety of word and document association techniques and of synonymous word and phrase groupings which serve to "clue" the selection of a subject heading are also being investigated by members of the Harvard group and guest investigators.

[1/] Major difficulties have to do with limitations both upon grammars and vocabularies so far tested and with ambiguities and the number of alternative parsings generated. See, for example, Bobrow, 1963 [68]. Kuno and Oettinger, 1963 [341] and Robinson, 1964 [502]. Bobrow provides a survey of syntactic analysis programs as of 1963, noting limitations or restrictions on each. He reports, for example, that available programs to compute word classes are not always correct in the class assignments made and that analysis systems are not complete unless they provide means for distinguishing between "meaningless strings and grammatical sentences whose meaning can be understood". He concludes: "Until a method of syntactic analysis provides, for example a means of mechanizing translation of natural language, processing of a natural language input to answer questions, or a means of generating some truly coherent discourse, the relative merit of each grammar will remain moot." ([68], p. 385) Robinson ([502], p. 12) says of sentences which can be parsed correctly, that they are: "Usually short sentences with no complicated embeddings of relative clauses and few participial or prepositional phrase modifiers. These include the basic sentences that most grammars are equipped to handle and that adult writers seldom produce."

Another partial approach to applying syntactic analysis techniques to automatic indexing is based upon syntactic word-class recognitions. Giuliano and his associates at Arthur D. Little, Inc., (1963 [230]), have investigated on a small-scale basis the use of the Kuno-Oettinger programs developed at Harvard for this purpose (Kuno and Oettinger, 1963 [340]). The broad program of information and language data processing research at System Development Corporation specifically includes investigations of structural patterns of sentences at the syntactic level and also of semantic factors such as the studies of polysemy and homographic ambiguity by Doyle, Wasser, and others. Borko reports:

> "...We...are analyzing actual written text for multiple meanings...The data for this study were drawn from the corpus of 618 psychological abstracts. Tabulations of frequency of paired and single word listings were used. A number of corpus-derived word frames have been prepared. Although this research is still in its early phase, we feel that we have made a good start on the problems of semantic analysis." 1/

In Czechoslovakia, at the Karlova Universita, both statistical and semantical methods for' automatic abstracting are reported as being under consideration. 2/

Other examples of proposals for the use of syntactic analysis techniques for the improvement of automatic indexing products include those of Spangler, Levery, Plath, Thorne, and Climenson and his colleagues at RCA, as well as the suggestions of those whose interests in automatic syntactic analysis have been primarily directed to problems of machine translation or more general problems of linguistic analysis. Hays, for example, although principally concerned with MT, indicates that the methods for determining phrase structures have obvious applications to the automatic determination of categories useful in the indexing of documents. 3/

An existing GE-225 computer program for KWIC-type indexing from both titles and abstracts at General Electric's Phoenix Laboratories is being extended to incorporate word analysis features taking into account both syntactic and semantic aspects of a given line or sentence of text. 4/ Levery provides an example of similar directions being explored in European research, more generally oriented toward linguistic considerations as such than to machine-derivable criteria (largely statistical to date), which seek to combine the benefits of both human and machine processes by way of automatic syntactic analyses. He claims, for example, that:

1/ Borko, 1962 [75], p. 6.

2/ National Science Foundation's CR&D report, No. 11 [430], p. 123.

3/ See Hays, 1961, [258], p. 13: "...Two broad problems on which work is just beginning at RAND: grammatic transformations and distributional semantics. The latter problems are especially important for automatic indexing, abstracting, and text searching." See also de Grolier, 1962 [152], p. 137.

4/ National Science Foundation's CR&D report No. 11 [430], p. 21.

> "... The study of the position of keywords in the text and the syntactical relationship which exists among them will show the way to automatic abstracting and the use of more sophisticated retrieval systems." [1]

Plath suggests that, given a computer program to perform the parsing and syntactic diagramming of a text sentence, the results can serve quite usefully to augment the selection criteria based initially on statistical techniques, such as word-frequency counting. He says, for example:

> "Another possible application of the outputs of the sentence diagramming program is their employment as an aid in language data processing for purposes of information retrieval, particularly in systems for automatic literature abstracting of the sort proposed by Luhn (1958). The feature of the tree diagrams which is pertinent here is that the main components of a clause, including subject, verb and object, always correspond to the 'main topics' in an outline, and are therefore located at the upper levels of the tree. When the words on these upper levels are considered apart from the lower-level structures which modify them, they often summarize the content of the sentence in a sort of 'newspaper headline' or 'telegraphic style'." [2]

The problems of multi-level selection, or screening, such that machine programs for selection of the most probably significant words, phrases, or sentences can be focussed upon the most probably content-relevatory areas of text, are treated here, as also by Salton, in the sense of a cutting-off at a given depth in the analyzed syntactic structure. [3] A potentially important contribution to the future prospects for automatic indexing, however, lies in the "discourse analysis" and "transformational linguistics" approach of Harris (1959 [254]), where condensations and concentrations of similarities and differences of topical interest may hopefully be achieved.

Harris himself suggested, at least as early as 1958, applications of his approach to both automatic indexing and abstracting. A goal of the analyses he has proposed is to identify 'kernels' of linguistic expression, having first, by various transformations such as from passive to active voice, brought together different ways of saying the same thing. He then suggests not only machine operations to normalize by application of his transformational rules but also to determine:

> "... Which kernels have the same centers in different relations (e.g., with different adjuncts), and other characterizing conditions. The results of this comparison would indicate whether a kernel is to be rejected or transformed into a section... of an adjoining kernel, or stored, and whether it is to be indexed, and perhaps whether it is to be included in the abstract." [4]

[1] Levery, 1963 [359], p. 236.

[2] Plath, 1962 [474], pp. 189-190.

[3] See also Thorne, 1962 [605], p.v: "The approach followed requires that the computer itself syntactically analyse input text in order to convert it into special form called FLEX, which preserves only that syntactic information which is useful for data retrieval purposes."

[4] Harris, 1958 [254], p. 949.

Certain difficulties are self-evident. Consider, for example, the admittedly hypothetical text which might refer in various places to the "dissolute, disreputable, illiterate, elder Lincoln" (underlining supplied) and which might be so processed by machine as to imply that Lincoln the son was, although also President of the United States, "dissolute," "disreputable," "illiterate," and "elder." These, however, are difficulties that plague almost any machine processing of natural language text.

Climenson, Hardwick, and Jacobson have explored some of the possibilities of the Harris approach in experimental computer programs for the RCA 501 (1961 [133]). Specific features of these programs include:

1. Establishment of the syntactic class or classes to which a given word can belong, by dictionary lookup.

2. Investigations of sentence structure and context in an attempt to resolve the homographic ambiguities involved when the same word may function either as a noun or a verb.

3. Isolation and marking of sentence segments, such as noun phrases, prepositional phrases, adverbial phrases, and verb phrases.

4. Identification and marking of segments -- clauses or degenerate clauses.

On a very preliminary basis, a limited set of word and phrase deletion rules were set up and several sample documents were processed against them, yielding reductions to about 35 percent of the original text. These results suggest that "syntactical filtering criteria" might be applied to the improvement of modified derivative indexing techniques, such as the word-frequency counting techniques, either by deleting syntactically insignificant parts of selected sentences, or by counting identical phrases rather than words. The investigators conclude, however, that:

> "A formal linguistic approach to the problems of natural language processing promises to yield results vital to the success of automatic indexing and data extraction. But the work required in such an approach will be quite arduous; a long-range man-machine effort will be required to formulate practical machine programs for indexing and abstracting." 1/

A final special case of linguistic data processing involving syntactic analysis is that of Langevin and Owens. They claim:

> "A critical review of the analysis work done on the Nuclear Test Ban Treaty by use of the Multiple Path Syntactic Analyzer demonstrates that such a device can, even at present, provide a powerful technique for the systematic discovery of ambiguities in treaties and other documents. Because the analyzer operates without bias from the overall context of the document, it may sometimes be possible for it to discover ambiguities that would easily escape a human reviewer who knows what the document is 'supposed to say'. " 2/

1/ Climenson et al, 1961 [133], p. 182.

2/ Langevin and Owens, 1963 [346], p. 26.

6.4 Probabilistic Indexing and Natural Language Text Searching

As in the case of automatic indexing proposals based upon automatic sentence extraction techniques, machine searching of full natural language text has been suggested as a basis for, at least, automatic derivative indexing. We have remarked previously that the machine use of complete text can only be considered to be "indexing" in a very special sense, that it is subject either to the non-availability of suitable corpora already in machine-usable form or to high costs of conversion to this form, and that too little is yet known of linguistic analysis and searching-selection strategies effectively applicable to natural language materials. Various examples of corroborating opinion, other than those previously cited, are as follows:

> "Machine searching is superb if it is known exactly how to describe the object of search, and if one could know how to choose from among many possible searching strategies. I doubt if any one is yet in this comfortable position with respect to machine searching of text." [1]

> "The most effective programs in automatic linguistic analysis have served only to illustrate how really complex is the structure of the language, and how far removed the present state of the art is from any system which might be useful in practice."[2]

> "The recognition of words involves only the matching of digital codes, but the recognition of an idea is a severe intellectual problem, the solution to which will probably never be exact. Nevertheless, this is the problem which must be attacked if accuracy is ever to be attained, or even approached, in using the text of information items as a basis for their recovery." [3]

Nevertheless, some of the work both in natural language text searching and in "probabilistic indexing" (where weights representing judgments as to degree of relevance of an indexing term to an item are used either in indexing or search), provide instructive insights into some of the problems of automatic indexing.

In the period 1958-1960, work at Ramo-Wooldridge resulted in the release or publication of provocative papers by Maron, Kuhns, and Ray on "probabilistic indexing" (1959 [398], 1960 [397]) and by Swanson on natural language text searching by computer (1960 [587, 582], 1963 [583]). Subsequent work along these lines has included further developments at Thompson Ramo-Wooldridge, the law statutes work at the Health Law Center at the University of Pittsburgh, and the experimental investigations of Eldridge and Dennis in a project jointly sponsored by the American Bar Foundation, IBM, and the Council on Library Resources.

[1] Doyle, 1959 [168], p. 2.

[2] Salton, 1962 [520], p. III,-1 through III-2.

[3] Doyle, 1959 [165], p. 12.

6.4.1 Probabilistic Indexing - Maron, Kuhns, and Ray

The work in the area of "probabilistic indexing" involves, as in the case of Stiles' statistical association factors, an assumption that there should be machine means available for the automatic elaboration of search requests in order that relevant documents not indexed by the precise terms of these requests may be retrieved. Given that measures of "closenesses" and "distances" between similar documents can be obtained, probabilistic weighting factors between index terms assigned to documents may be made explicit. More generally, however, the notion of probabilistic indexing is based upon the assignment of weights that provide a numerical evaluation of the probable relevance of index terms to a particular document, and of the relative importance of the various terms used in a search request. Maron and Kuhns (1963 [397]) thus consider the following variables important in the formulation and following out of search strategies:

1. Input- both the terms of the request and the weights assigned to them.

2. A probabilistic matrix giving dissimilarity measures between documents, significance measures for index terms, and closeness measures between index terms.

3. A priori probability distribution data.

4. Output- a class of retrieved documents ranked in order of their "computed relevance numbers" and an indication of the number of documents involved in the class.

5. Search parameter controls, such as the number of documents desired.

6. Search prescription renegotiation involving amplification of the request by adding terms "close" to the ones in the original request and the selection of additional documents following distance criteria for the collection.[1/]

Experiments have been reported for 40 requests run against 110 articles taken from Science News Letter. Without search renegotiation, the "answer" document was retrieved in only 27 of the 40 tests. Three alternative methods of request elaboration were then tried. First, additional terms most strongly implied, statistically, by the terms in the request were used. Secondly, those terms were added which most strongly imply, again in a statistical sense, each of the given request terms. Thirdly, coefficients of association between index terms were used. Results are reported as follows:

"(1) Using the method of request elaboration via forward conditional probabilities between index tags, we retrieved the correct answer document in 32 cases out of the 40.

(2) Elaborating the requests via the inverse conditional probability heuristic, we retrieved the correct document in 33 of the 40 cases.

(3) Using the coefficient of associationto obtain the elaborated request we obtained success in 33 cases of the 40.

[1/] Maron and Kuhns, 1960 [397], pp. 230-231.

> "Thus we see that the automatic elaboration of a request does, in fact, catch relevant documents that were not retrieved by the original request." [1]

6.4.2 Natural Language Text Searching - Swanson

The work in automatic indexing and related research directed by Swanson at Ramo Wooldridge Corporation has included "indexing at the time of search" in natural language text searching, (1960 [582, 587], 1963 [583]), the previously mentioned studies of machine-like indexing by people (Montgomery and Swanson, 1962 [421]), and automatic assignment indexing using pre-selected lists of clue words, (Swanson, 1963 [580]). The last of these three major areas of investigation is the one of the greatest interest in this present study, but the earlier experiments in machine searching of natural language texts warrant some discussion. In his reports on this text searching project, Swanson has specifically claimed that the methods for transforming search questions can serve as the basis for an automatic indexing method. Thus:

> "...A technique for automatic indexing can be derived immediately from a text searching technique...it is necessary only to so organize the machine procedures that those operations of text reduction or reorganization common to all searches are performed only once and prior to searching in order to create directly an automatic indexing procedure." [2]

Swanson has also claimed that if automatic searching of full text is not feasible, then automatic indexing is not feasible, the one being prerequisite to the other. For example:

> "Clearly, if a computer technique for search and retrieval from the full text of a collection of documents cannot be developed, then it is unthinkable that matters could be improved by using the machine to operate on just part of the information (a 'condensed representation') -- that is, on an automatically produced index. This line of argument demonstrates persuasively that the development of techniques for automatic full-text search and retrieval is a prerequisite to automatic indexing. It is equally clear that a technique for automatic indexing can be derived immediately from a text-searching technique, and thus that the two processes involve conceptually equivalent problems." [3]

In the actual text searching experiments, a model "library" consisting of 100 short articles in the field of nuclear physics was set up in machine-usable form. These articles were also studied by subject specialists who rated the relevance of each paper to each of 50 questions, and assigned weighting factors representing the degree of judged relevance. A second group of people, who knew only that the papers were in the field of nuclear

[1] Ibid, p. 240.

[2] Swanson, 1960 [582], p. 6.

[3] Swanson, 1960 [587], p. 1100.

physics, then transformed the 50 questions into search prescriptions using three different methods. The first method for the development of the search instructions was to choose appropriate index entries from a subject heading list tailored to the contents of the sample library. Search was then made manually against a card catalog which recorded the results of manual indexing of the same 100 articles to the entries of this list.

The second method of search prescription tested involved the specification of combinations of words and phrases likely to be found in any paper which would in fact be relevant to the search question. The third method involved modification of the second by the use of a thesaurus-type glossary which suggested various alternative terms. Both the latter two types of search instructions were fed to a computer program which carried out searches against the natural language text consisting of 250,000 words from the original articles.

The results were then evaluated in terms of ratings of relevance made by the physicists who had analyzed the papers. Retrieval effectiveness was not high: "...in no case did the average amount of relevant material ... retrieval (taken over 50 questions) exceed 42 per cent of that which was judged ... to be present in the library." [1/] However, the results were indicative of the superiority of the machine methods to the manual catalog search.[2/] For this library in particular, in the case of "source documents" (the articles from which the search questions were taken), only 38 percent of the relevant papers were located by the manual search, whereas 68 percent of the relevant items were retrieved by machine search of the text for specified words and phrases in various "and" and "or" combinations. Machine search based on search instructions that had been developed with the assistance of the thesaurus-glossary yielded 86 percent of the relevant source item documents.

6.4.3 Full Text Searching - Legal Literature

"The retriever of documents may be satisfied with a sample of descriptors that represent the contents; the fact retriever or the question answerer must often have access to every word in the text". [3/] The objective of fact retrieval is a major goal in the experimentation that is being carried forward in the field of natural language text searching of legal material, especially the texts of statutes of the State and Federal Governments. The most extensive program to date is that of Horty and his colleagues at the University of Pittsburgh Health Law Center (1960 [277], 1961 [276, 309], 1962 [196, 278], 1963 [24, 280]).

Wilson at the Southwestern Legal Foundation is experimenting with a modified version of the Horty-Pittsburgh System for legal cases dealing with arbitration in five of

[1/] Swanson, 1960 [582], p. 25.

[2/] Ibid, p. 1: "On the whole, retrieval effectiveness was rather poor, yet machine search of the text of the model library was significantly better than was human searching of the subject heading index."

[3/] Simmons and McConlogue, 1962 [555], p. 3.

the southwestern states.[1] A joint American Bar Foundation--IBM research program has been established to explore both text searching without prior indexing and automatic indexing techniques (Eldridge and Dennis, 1962 [183], 1963 [182]).

In the Horty-Pittsburgh System, approximately 6,000,000 words of text have been converted via Flexowriter to magnetic tape. An exclusion dictionary of 100 words is used to eliminate the most common words and a word-concordance is prepared, resulting in word-occurrence location indicia by position in sentence, paragraph and section of the statute. In searching, the user has available to him the alphabetized list of approximately 17,000 different words and it is up to him to think of the words and synonyms most likely to occur in statute sections likely to be the ones he seeks. Several search logics are available. One provides that at least one of a group of alternate words must appear; another requires that at least one from two or more groups must appear in the same sentence. Intra-sentence distance criteria are also utilized: "If the phrase 'born out of wedlock' is sought, the operator... requires that the word 'wedlock' appear in the same sentence, no more than three words after 'born'." [2]

Obviously, for the same question the searcher would also have to specify synonymous words and phrases--"illegitimate children", "illegitimate births", "unwed mothers", "unmarried mothers", "illegitimacy", "bastardy", and so on. The reported success of the system is apparently due in large part to the ingenuity of the searchers in specifying the expressions and synonyms most likely to be used. Hughes comments as follows:

> "It should be noted that this system will be most efficient only when the users are thoroughly familiar with the linguistic style of the source material and search is made on words known to occur in the appropriate statutes". [3]

6.5 Other Examples of Related Research in Linguistic Data Processing

Since, as Garvin has emphasized, "All areas of linguistic information processing are concerned with the treatment of the content, rather than merely the form, of documents composed in a natural language," [4] much of the research in linguistic data processing is potentially applicable to both the development and the improvement of automatic indexing techniques. Thus developments in automatic content analysis, in psycholinguistics, in question-answering systems, may eventually find application to mechanized indexing systems.

[1] Eldridge and Dennis, 1964 [182], p. 90; Wilson, 1962 [645].

[2] Horty, 1962 [278], pp. 59-60.

[3] Hughes, 1962 [284], p. IV-6 to IV-8.

In terms of our present concern, however, we shall select only a few examples. "By automatic content analysis is meant the use of computer programs to detect or select content themes in a sentence-by-sentence scanning of text or verbal protocols".[1] The interest of psychologists in machine techniques to assist in the analysis of linguistically-given materials, as in propaganda analysis, probably precedes at least in sophistication if not by date, that of documentalists or of machine specialists interested in library and information problems.[2]

The "General Inquirer" program developed by Stone et al,[3] is an example of question-answering techniques based upon selective extractions from natural language text. It involves the use of a master vocabulary consisting of words previously selected by an investigator as being likely to be content-indicative in a body of material to be processed, together with his pre-established indications of the categories he expects their occurrence should predict. It is to be noted that this is a custom-tailored set of categories and of clue-word lists associated with each, manually pre-established. Text is now processed in such way that each word is looked up and, if it appears in the master vocabulary, it is tagged with identifiers of the categories for which it is presumably predictive. A subsequent "Tag Tally" routine then counts the tag frequencies to determine for which categories the input material has high or low scores, and these in turn can be compared with expected norms.

This type of program has been applied to such varied materials as suicide notes, folk tales from different cultures, reports of field workers, recordings of group discussions as in supervisory-leadership training sessions, and protocols for various psychological tests.[4] Interesting variations developed by Jaffe and others[5] involve the use of non-verbal as well as verbal clues as content-indicators, specifically, time-sequence patterns recorded along with the words spoken in client-therapist sessions. At the meeting of the Association for Computational Linguistics and Machine Translation held in Denver, August, 1963, Jaffe reported findings indicative of positive correlation between the structure of temporal and lexical patterns in dialogue and suggested applications to automatic abstracting or indexing by the use of the time-sequence patterns as clues to high information-value areas.

[1] Ford, Jr., 1963 [498], p. 3.

[2] See, for example, Jaffe 1952 [297], Hart and Bach, 1959 [256], Pool, 1959 [475], the latter covering the proceedings of a conference held in 1955.

[3] Stone and Hunt, 1963 [576]; Stone et al, 1962 [575].

[4] See Ford, 1963 [498], p. 8.

[5] See for example, Cassotta, et al, 1964 [104]; Jaffe, [294] to [297].

Hughes provides, as of September, 1962 ([284]), a critical review of several experimental and proposed question-answering systems using natural language statements and natural language queries, including "BASEBALL", [1] "SAD SAM" [2] and the "Proto Synthex" investigations of System Development Corporation. [3] Later developments on the Synthex (synthesis of complex verbal material) project at SDC have included a variation on a natural language text searching program where ordinary text input is run against an exclusion list and a table is set up to tally the substantive words remaining. Words with the same roots or previously having been identified as synonymous are cross-referenced. A complete index results, with document location identifier tags for the word occurrences down to the single sentence level. This index can be used subsequently to locate regions of text (volume, chapter, paragraph, and sentence) where answers responsive to input questions are likely to be found.

It is proposed that the Synthex system eventually should incorporate analyses of syntactic and semantic relationships in the linguistic expressions of both queries and text. Of future interest in the extension of such considerations to automatic indexing and abstracting are the following comments:

> "The results of several early experiments within the project, coupled with the findings of other language researchers, led to the following conclusions about meaning and grammatical structure in English text:
>
> 1. The degree of synonymity in meaning between any two English words can be measured quantitatively with a synonym dictionary and relatively simple scoring procedures.
>
> 2. The difference in meaning between two sentences of identical syntactic structure can be expressed quantitatively as a function of synonymity of their words..." [4]

It is also of interest to note that although the "indexer" program of the Synthex system provides cross-referencing between, for example, "whales" and "whaling" or "England" and "Great Britain", the investigators admit that: "naturally it falls short of such complicated cross-referencing as 'mouse-animal' 'Jones person' and other concept recognitions." [5] However, concept recognitions based upon both *a priori* and

[1] See also Green et al, 1961 [238].

[2] See also Lindsay, 1960 [363].

[3] See also Klein and Simmons, 1961 [325]; Simmons et al [552]to [555].

[4] System Development Corporation, 1962 [590].

[5] Simmons and McConlogue, 1962 [555], p. 70.

a posteriori associations are at least foreshadowed in a small-scale model of attribute-words and proper names, together with prespecified relationships between them; [1] in Olney's recent work at SDC exploring the possibilities for use of cognitive concepts as bases for establishing association between documents, [2] and by Kochen's work on machine' inference and concept processing. [3]

A final example of potentially related research in the area of content analysis is therefore the work of Kochen, Abraham, Wong and others at IBM's Thomas J. Watson Laboratories (1962 [329]). While concerned principally with adaptive organization and processing of stored factual statements and the possibilities for machine formulation of "hypotheses" about these and additional facts, some consideration has been given to sampling procedures applicable to determination of similarity which might be used for document clustering and to the possibilities for dynamic clustering for retrieval based upon a specific individual query. [4] In the proposed AMNIP (Adaptive Man-machine Non-Arithmetical Information Processing) system„ there is no attempt at either automatic indexing or automatic abstracting. [5] Instead, formal statements are made about named "things" and their attributes. The sharing of common attributes then serves as a basis for relating items which are similar and for grouping them together in the system memory. It is assumed that the organization of the stored statements changes dynamically with new data inputs and user feedback in question-answering routines.

Where the named items are names of documents or of index terms, a number of documentation applications can be considered. Where the items are document names and the formal predicate is "cites", the system provides a procedure for production and use of citation indexes. [6] Where the items are index terms or subject headings and the predicates are "is used synonymously with" or "is subsumed under", machine construction of a growing thesaurus based on use is suggested. [7] The common attribute

[1] See Stevens, 1960 [568]; see also Herner, 1962 [266], p. 5.

[2] See Borko, 1962 [75], p. 5: "Instead of defining meaning in terms of synonyms... it is defined in terms of the entities referred to by the word in context. A chair is thus described as belonging to a class defined by a given list of properties... Analysis yields an interpretation of the sentence as an assertion that certain relationships hold between the specified referent classes. The cognitive content of the sentence is a function of this assertion plus the information about these referent classes which has previously been stored in memory."

[3] Kochen et al, 1962 [329] .

[4] Ibid, Appendix by C. T. Abraham, pp. 20-65.

[5] Kochen et al, 1962 [328], p. 45.

[6] Ibid, p. 37.

[7] Ibid, p. 37.

matching program, applied to logical similarities of texts related as by having various assigned descriptors or citations in common, might provide a basis for generating document surrogates by representing each text in a related group of texts with the words or sentences these texts have in common. [1]

In the case of man-machine interaction during search, it is suggested that the user should indicate the names of selected documentary items which are of particular interest, then:

> "The machine forms an 'hypothesis' about the subset of articles likely to be of interest. It does this by examining all recorded statements common to the ones selected but not to the rejected ones. The weight of different attributes and degree of interest is taken into account. The machine may display this hypothesis or another random sample of titles consistent with it, or both." [2]

6.6 Machine Assistance in Translations of Subject Content Indications to Special Search and Retrieval Language

There are, also, in the areas of directly and indirectly related research, certain programs of research, development, and experimentation which include investigations of possibilities for using machines to assist in the "translation" of textual languages into special intermediate or "documentary" languages. Doyle's use of the inclusion list principle to extract specified content-indicative words and to encode them in his "bigram" index was an early but relatively trivial example. [3] The work of Williams and her associates, at Itek and elsewhere, [4] has involved the objectives of determining which of the subject-revealing implications of titles, abstracts and, if necessary, full text, are susceptible to machine detection and manipulation such that the implied as well as the explicit assertions made in a document may be incorporated in a formalized language for retrieval.

While Williams, Barnes, Cardin and Levy, and others, have so far approached such tasks primarily from the standpoint of human analytic judgments, Coyaud (1963 [143]) has discussed at least preliminary work looking toward the automation of the analysis of natural language texts for purposes of encoding and organization of the terms and relationships to be used in the "documentation language" known as "SYNTOL" (Syntagmatic Organization of Language), this work has used a corpus based on bibliographic abstracts from the Bulletin signalétique of the Centre National de la Recherche Scientifique, Psychophysiology Section, for the period 1958-1960. Notwithstanding such difficulties as determining rules for proper subdivisions of text, reduction of synonyms, resolution of lexical and syntactic ambiguities, and the fact that some words are always,

[1] Kochen et al, 1962 [329], p. 2.

[2] Kochen et al, 1962 [328], p. 7.

[3] Doyle, 1959 [168]. See also p. 123 of this report.

[4] See, for example, T. M. Williams, R. F. Barnes, Jr., J. W. Kuipers, various references.

but some never, used in SYNTOL itself, he reports that both substantives and textual expressions indicative of certain specific SYNTOL relations can be unambiguously identified. Contextual clues are used: for example, if the word "homme" occurs it is translated as "sexe masculin" if "femme" also occurs, as "etre humain" if "animal" is also mentioned, and as "sujet experimental" otherwise.

Melton and her associates at the Center for Documentation and Communication Research, Western Reserve, have also been investigating machine processing of input text with a view to the automatic selection and manipulation of clue words and relationships between them for information retrieval purposes. Their material consists of abstracts from the metallurgical section of Chemical Abstracts. From sample abstracts, a lexicon is developed which involves classification of words into those that are significant from a metallurgical point of view;those that name materials, compounds, environments; those denoting processes; those denoting characteristics of materials; prepositions; those which will not operate in the analysis of the text, and the like.

On the basis of analysis of a number of sentences from the sample text, rules for combination and selection of specified words in specified relationships can be set up. These rules are designed to identify sentence types which:

(1) Describe performance of a process on a material.

(2) Discuss a material in terms of properties, components, form, or environment.

(3) Describe a process without reference to specific materials.

(4) Discuss metallurgical properties without reference to specific materials.

(5) Discuss two or more materials, properties or processes.

(6) Describe a causal relationship between two properties.

(7) Give a comparison of materials.

(8) Contain no words of interest in the system.

Computer programs to explore the possibilities for automatic analyses of the kind developed manually for the sample abstracts will be written with the objective of finding an effective compromise between mere word identification and total linguistic analysis. Melton says:

> "If one considers this method of analysis from the point of view of the linguist, he can immediately describe many grammatical constructions, which will prevent the meaningful reduction of these sentences. It is not known at this time how often such sentences will appear in the corpus of this investigation. Nor is it known how adversely such failure would affect the retrieval of the information in these sentences. The answers to these questions will be available only after a large sample has been analyzed and put to an extensive retrieval test. At its most successful the project will achieve an automatic processing of metallurgical text which will permit retrieval of the type of information which can be stated in its own terms with a tolerable amount of inappropriate selections. Should this goal be unattainable, the project will have generated a file of abstracts automatically searchable on the word level

or somewhat beyond. For the benefit of other research, it will also have produced tapes of the true text of a large sample of natural-language abstracts and a lexicon containing all the words of a corpus of current scientific literature." [1]

6.7 Example of a Proposed Indexing-System Utilizing Related Research Techniques

In addition to the automatic assignment indexing and automatic classification techniques for which experimental results have been reported, several other techniques and programs have been proposed. One is the joint American Bar Association-IBM research program (Eldridge and Dennis, 1963 [182]), for which discussion has been deferred because of its proposed use of several of the research techniques covered previously in this section. The experimental corpus will consist of the full text of approximately 5,000 legal case reports taken chronologically from the Northeastern Reporter. Approximately half of this material will be processed to obtain word frequency counts. The frequencies will then be used to prepare for each different word an estimate of the skewness of its distribution in the collection. The investigators will then personally inspect the word list as ordered by skewness to divide it into "non-informing" (Type I words, or an exclusion list) and "informing" (Type II words, or an inclusion list) at some appropriate cutting point. Then, for each document, a list will be prepared of its "informing" (Type II) words, maintaining order within the document. For each pair of such words, statistical association factors will be computed. Eldridge and Dennis describe other aspects of their proposed technique, in part, as follows:

> "For each document in the body of 2,500 cases, a list will be prepared of its Type II words, maintaining their original order within the document ... For each Type II word an 'association factor' will be calculated for every other Type II word with which it appears in any one document by compiling the probability that Word A would appear this close to Word B this number of tries over the entire file, if the Type II words were distributed at random. (This amounts to borrowing Stiles' idea of the association factor, but implementing it with a numerical method which takes into account nearness of the words within the document as well as the fact that they both occur in the same document.) Since the factors are probabilities, they will be numbers between zero and one ... These numbers will be used to estimate the distances between words in index-word space.
>
> "The next step is to construct from the information about distances between pairs of words an index-word space in which every word is at the correct (or approximately correct) distance from every other word in the system with which it exhibits association. The result of this operation can be visualized schematically as a sort of grid in which every word can be placed in its appropriate position by assigning it a set of coordinates."

[1] Melton, et al 1963 [414], pp. 14-15.

> "Indexing of the remaining cases in the experiment will be performed by machine from full text, using the Type I list of discard words and the Type II list to prepare an analysis of the frequencies related to index-word space. Instead of selecting specific words as indexing terms, concepts will be selected (statistically) as volumes in index-word space. A rough physical analogy to this process would be to toss pennies at the previously mentioned grid so that, for every Type II word in the source document, a penny lands at its proper slot on the grid. Where the pennies heap up in a pile, you have a concept."
>
> "Searching will be carried out essentially by indexing a question presented narratively, determining the concept volumes that represent the question, and searching those volumes in document space for the relevant document numbers. Since the 'edges' of the concept volumes are determined statistically, output can be listed in order of probable relevance; as an option the question could be accompanied by a request that 'at least 100 references be supplied', in which case the concept boundaries would be adjusted to provide that number." 1/

It will thus be noted that the proposed indexing and search program begins on a derivative basis to establish for one-half the experimental material the significant words, next combines word frequency with significant word distance data to derive probabilistic association factors between words, then develops clusters, and finally indexes the items in terms of the clusters rather than words so as to provide assignment rather than extraction of index terms.

7. PROBLEMS OF EVALUATION

We have noted, in the introduction to this report, that several fundamental and highly controversial questions can be raised with respect to the feasibility and evaluation of any automatic indexing scheme and with respect to the evaluation of any indexing systems whatsoever. Yet if automatic indexing procedures are to be based upon previous human indexing or if their results are to be compared with human results, then the questions of the quality, the reliability and the consistency of human indexing are crucial ones indeed. Thus, Solomonoff warns:

> "The finding of exact languages for retrieval is also made less likely, in view of the fact that the categorizations of documents that are presented to the machine as a training sequence will not be performed altogether consistently by the human cataloger." 2/

Montgomery and Swanson ask whether human indexers are in fact self-consistent and consistent with each other, and they suggest:

1/ Eldridge and Dennis, 1963 [182], pp. 97-99.

2/ Solomonoff, 1959 [562], pp. 9-10.