The proposed "Multilindex" system is also based on micro-thesauri or small vocabularies designed, by human analysis, for clue-indications to a relatively narrow subject field, together with potential syntactic-semantic role indications built into the dictionary, again by extensive human analysis, following the approaches previously taken by A. L. (Lukjanow) Loewenthal in her suggestions for solutions to problems of mechanized translation. An unpublished proposal-type brochure describing the system was available as of December 1963.[1] As of that date, also, demonstration printouts were available from an IBM 1401 Fortran program, illustrating an index compiled from abstract-text input and a 1,200-word dictionary for documents in the field of space antenna tracking radar. [2] A repetoire of 350 "concepts" or indexing terms was involved, with an average of 10 assigned to 22 test documents, many of these assigned terms being identical to words occurring in either the title or the text of the abstract of the item.

Slamecka and Zunde have investigated the extent to which the "notations-of-content" in the system developed by Documentation, Inc. for NASA's STAR might be derived by machine techniques from the text of the abstracts with enough normalization-standardization via inclusion dictionary lookup to qualify as an assignment indexing technique. These workers claim:

> "This preliminary investigation indicates the possibility of using the computer to index documents adequately for machine retrieval by matching their abstracts against an authoritative subject-heading authority ... The inconsistency inherent in human indexing can be eliminated as the number of terms derived from any one abstract will always be the same. The abstract and its automatically derived set of index terms will always be equivalent..."[3]

A final example of other approaches to automatic assignment indexing research, not yet reported in the open literature, is an NIH sponsored project at Goodyear Aerospace, in cooperation with the Universities of Minnesota and Rochester and Western Reserve University, looking toward an automatic classification procedure based on word coocurrences for a set consiting of 100 four-to-five page documents in the field of diabetes literature. Programs for statistical analyses of the full text of these documents, all of which have previously been processed for the manual W. R. U. "telegraphic" abstracting system, are being developed. [4]

5. AUTOMATIC CLASSIFICATION AND CATEGORIZATION

In all the experimental work, to date, that has been directed toward the use of computers and other machine-like techniques for the automatic indexing of documents, a

[1] "Description of MULTILINDEX. A mechanized system for indexing documents, storing information, retrieving information", P.S. Shane, Dec. 4, 1963, Information Systems, Inc., 7720 Wisconsin Avenue, Bethesda, Maryland.

[2] Private communications, A.L. Loewenthal and P.S. Shane, Dec. 11, 1963.

[3] Slamecka and Zunde, 1963, [561], pp. 139-140.

[4] E. Tuttle, private communication, Oct. 30, 1963.

dichotomy can be observed. There is, on the one hand, a spate of examples of automatic derivative indexing where words used by the author himself or by human analysis are sorted and arranged, by machine, to provide index listings, announcement bulletins, and current awareness distribution notices. There are also, on the other hand, at least a few instances of investigations where the machine assigns category labels, indexing terms, or "heads" and "headings" from a classification schedule, to new items.

In general, as Needham [1/] points out, proposed automatic assignment indexing procedures can be investigated with reference to a previously existing index term vocabulary, an existing classification system or schedule, or to specially designed vocabularies and subject heading lists. On the other hand, if it is not known how well existing systems do in fact characterize documents and if it is not known whether all pertinent properties of the documents have been consistently identified, then it may be preferable to develop methods for assigning documents to the appropriate class in a classification system which is itself set up automatically. [2/] Needham also suggests still a third possibility: that of setting up automatically a classification within which the subsequent classifying of documents is done by hand.

The principal experimental results, to date, of attempts to achieve automatic classification of documentary items, especially in the sense of machine-generated groupings or categorizations of such items, have been those of applying techniques of "clumping", [3/] factor analysis, and "latent class analysis". [4/] We shall briefly consider below some typical investigations into automatic classification or categorization procedures that have already had, or may have, applicability in automatic index ing techniques.

In the late 1950's, Tanimoto undertook theoretical studies of mathematical approaches to problems of classification and prediction with special reference to matrix manipulations of sets of attributes of items to be classified. [5/] He also investigated

[1/] Needham, 1963, [432], p. 1.

[2/] Ibid, p. 1-2: "If we are to assign a document to a class automatically, we must have a) a list of facts about the classes which will make ascription possible: b) an algorithm, usually some sort of matching algorithm, to tell us which class best suits a document. Given a classification like the U. D. C., it is not at all obvious that a) and b) exist, or even, if they can be found. a) and b) imply a degree of uniformity about the classification which may just not be there."

[3/] That is, the clustering of objects that are in some sense similar because they share certain attributes or properties, even if, and especially when, the identity of cluster-producing common properties is not known in advance.

[4/] Compare Doyle, 1963 [162], p. 13; "There are other statistical techniques besides factor analysis whose output is document clusters, such as latent class analysis and clump theory, and there is a surprising increase in research in this kind of analysis just within the last two years."

[5/] Tanimoto, 1958 [593], 1961 [594]. See also Borko, 1963 [76], pp. 4-5: "In 1958, Tanimoto published a theoretical paper on the applications of mathematics to the problems of classification and prediction. Specifically, he pointed out how the problems of classification can be formulated in terms of sets of attributes and manipulated as matrix functions."

theoretical aspects of automatic indexing and sentence extraction involving co-occurrences of words. While Tanimoto's studies with respect to linguistic information processing for classification purposes have apparently been limited to the theoretical considerations, similar concepts of probabilistic, computational, and matrix manipulative operations to derive and use coefficients of correlation of associations between such attributes as words occurring in text or the index terms assigned to documents are involved in the factor analysis and theory of clumps techniques as applied in actual experiments in documentary classification.

5.1 Factor Analysis

The factor analysis technique which seeks to derive from word associations in representative documents an automatically generated classification schedule for use in actual indexing experiments has previously been mentioned. [1] Reasons suggested for its use in research at SDC have been reported as follows:

> "The development of automatic procedures for purposes of classification and abstracting requires the identification and specification of attributes of words or passages so that the relevancy of topics or content can be determined. Automatic procedures to detect such attributes may be based on a number of characteristics of the text: word frequencies, syntactical information, semantic information and pragmatic contextual clues. Currently, word frequency information can be generated and manipulated by automatic procedures, whereas the other attributes are not as readily handled this way. However, a correlation matrix of content words becomes very unwieldy because of its size and the complexity of relationships. For this reason, factor analysis is used to identify clusters of relationships. Current work concentrates primarily on determining the usefulness of factors identified in this way as classification and indexing schemes." [2]

As noted above, Borko and Bernick (1961 [73], 1962 [77], 1963 [78]) have applied this technique to abstracts drawn from psychological literature and to the same computer literature abstracts as had been used by Maron, (1961 [395]). This technique had also been investigated in the studies looking toward information retrieval classification and grouping undertaken at the Cambridge Language Research Unit from about 1957 onward. However, certain apparent limitations of the factor analysis approach led Parker-Rhodes and Needham to the alternative of the "theory of clumps" (1960 [465], 1961 [435, 464]). Parker-Rhodes gives the rationale, and some of the distinctions between the two techniques, as follows:

> "It has been assumed that statistical methods could be applied to the data in such a way as to reveal any objectively existing classes which may be there. The general

[1] Pp. 94-97 of this report.

[2] System Development Corporation, 1962 [590], p. 15.

> name for the techniques evolved in this way is factor analysis. Insofar as it is practically applicable this technique has worked well enough; but...it has two limitations (a) that some classification problems are outside its scope, and (b) that it is not susceptible (at least as hitherto conceived) of adaptation computationally to the study of really large universes..." [1]
>
> "...The procedure of factor analysis first finds certain clumps, but then, as output, it gives us vectors relating the descriptors of the universe to the clumps found...
>
> "In most cases, factor analysis is used (especially in psychology) to debug the descriptor space; more conventionally put, to eliminate those tests (descriptors) which have an equivocal membership in several factors (Clumps) in favor of those which, having more definite allegiances, convey more information of the kind which the analysis suggests as valuable. It is thus only related to the classification of the universe at one remove; the classification it suggests is a simple categorical classification defined by the descriptors suggested as the most valuable...
>
> "The descriptive array of a universe is a table giving the applicability or inapplicability of each descriptor to each element. To classify the elements of the universe, we calculate for every pair of elements a similarity as a function of the corresponding rows of the descriptive array, and then regard the similarity matrix as a sufficient description of the universe. In factor analysis, on the contrary, we start with the matrix of correlations between the descriptors, each being a function of a pair of columns of the descriptive array..." [2]

Other investigators who have considered factor analysis techniques for possible applications to automatic indexing, automatic categorization of items in a collection of items, or search prescription renegotiation in a mechanized selection and retrieval system include Stiles (1962 [573]), Doyle (1963 [162]), and Hammond (1962 [251]).

Stiles, whose principal experimental results relate rather to the use of statistical associations between terms manually assigned to documents for search prescription formulation and renegotiation than to automatic indexing procedures as such, [3] has also considered both automatic indexing and automatic classification approaches. Specifically, he has made at least preliminary investigations of the factor analysis technique independently developed for similar purposes by Borko. For a large collection of 105,000 items, the statistics of co-occurrence of indexing terms were in some cases not as precise as desired because the same terms were used in different senses for different items in the collection.

[1] Note that Borko himself confirms this limitation as recently as November 1963, in stating, of the CLRU work on clumps: "However, even now these techniques have been applied to a 346x346 matrix which is beyong the capabilities of presently available factor analysis programs." (1963 [76], p. 8).

[2] Parker-Rhodes, 1961, [464], pp. 3-6.

[3] This principal concern is discussed below with reference to potentially related research, pp. 119-122 of this report.

The possibilities of using factor analysis to sort out the different meanings were therefore explored. [1] Using an IBM 704 program, the centroid method of factor analysis was applied to a matrix of correlation coefficients of terms that had co-occurred significantly with the term "exposure". Three factors were derived, one generally relating to the corrosive effects of exposure, another to "exposure" in the sense of photographic exposure, and the third dealing with both exposure-to-weather and exposure-to-radiation. Although the results were considered quite satisfactory, more extensive experimentation and use did not appear feasible because of computer matrix manipulation limitations.

Doyle notes, in particular, that factor analysis might be used to give well-defined clusters separated one from another by clear boundaries rather than the less precise clusters found by most document grouping techniques. He emphasizes, however, that "its success in doing so of course, depends on the well-defined clusters actually being present in the data". [2] He suggests that a combination of factor analysis and human editing to select items most typical of statistically derived categories could be valuable in such applications as the sorting of Congressional mail or the identification of trends in political or military intelligence materials free from the personal biases of an analyst.

Hammond and his Datatrol associates who have worked on an application of the Stiles association factor technique for search question negotiation to legal literature have also considered the potentialities of factor analysis. Thus they report:

> "... The present association factor gives the relationship of one term to another. A factor analysis study would allow us to determine the relationship of a single term to a group of terms. From this we could learn how terms cluster when related to the same concept." [3]

5.2 The Theory of Clumps

It is assumed, in the work on the theory of clumps, that we have a population of objects or items among which at least some classes or groupings do objectively exist, but that we do not have any bases for precisely determining class membership requirements. There may, therefore, be many possible ways of grouping and many possible definitions of clumps. On the other hand, such diverse definitions must conform to the extent of some similarities of membership in the clumps that they define if in fact they do define any of the existing classes. Assuming further that we are given information about properties ascribable to various members of the population, it is theorized that useful clumps can be discovered by investigating similarity connections between pairs of items, such as the number of co-occurrences of specific properties. Thereafter, only these similarity connections are considered, and the connection matrix is used as the basis for trial partitions of the population into various possible subsets.

[1] Stiles, 1962 [573], pp. 10-12.

[2] Doyle, 1963 [162], p. 12.

[3] Hammond, et al, 1962 [251], p. 17.

In early work on clump definition, Kuhns of Ramo-Wooldridge [1] proposed the use of a threshold value such that if a subset is a clump every pair of members in it has a connection strength equal to or greater than the threshold value and no member of the subset's complement has connections of more than threshold value to the members of the subset. In the more extensive investigations carried out by Parker-Rhodes and Needham (1960 [465], 1961 [434, 435, 464]), other clump definitions have been explored and specifically that of the "GR-Clump". This is defined as a subset of the universe such that all its members have a positive (or zero) bias to the subset and all non-members have a negative bias to it, where bias is defined as the excess (positive or negative) of the total connections of a member of the population to the members of the subset over its total connections to the members of the subset's complement, following the convention that the connection of the element to itself is taken as zero.

An iterative procedure for discovering GR-clumps can now be followed. This is based on an arbitrary initial partition of the given universe of elements into a subset and its complement. Then, since each element has a bias toward both the subset and its complement, differing only in sign, the biases of each element are computed. If the bias of a particular element is positive with respect to the subset, it is transferred to the subset if it is not already a member of it, and conversely if its bias is negative, it is transferred to the subset's complement if it is not already there. Each time a transfer is made, the biases are recomputed and the process is repeated until for a complete scan of all elements no further transfers can be made. The result is a GR-clump even though it may have no members or may contain all the elements of the universe. In such case, a further partition is made and the procedures are re-applied.

These GR-clump finding procedures have been applied to such diverse collections of items to be classified as archaeological artefacts and patients' symptoms as related to specific disease diagnosis. In the latter case, groupings were obtained that corresponded satisfactorily to certain specific disease syndromes, but no group was found corresponding to Hodgkin's disease where a great variety of symptoms typically occur. Needham comments: "I can scarcely conceive of a clump definition that would be likely to group these patients; I am unsure whether this is a reflection on clump theory or on Hodgkin's disease." [2]

In applications more directly related to documentation, some investigations have been made of the use of co-occurrence coefficients of index terms assigned to documents in order to form a connection matrix from which clumps were then derived (Needham, 1963 [431]). These experiments covered 342 terms occurring more than once in the index-term sets assigned to several hundred documents in the general subject field of machine translation. Computation of the matrix required 20 minutes of computer time and the 40 clumps found took 6-8 minutes each to find. Needham reports on the results as follows:

[1] See Kuhns, 1959 [336], and Needham, 1961 [435], pp. 20-21.

[2] Needham, 1961 [435], p. 46.

> "Evaluation of the results was unexpectedly difficult. The acid test is presumably the efficiency of the retrieval system embodying the grouping given by the program; but the efficiency of retrieval systems cannot be easily measured. An apparently simpler test would be to see if the clumps were intuitively satisfactory, i. e., were groupings that a classifier in his right mind could have made. This also was unsatisfactory because the groups are mostly rather large, larger in fact than classifiers ordinarily make, and were thus very difficult to judge. The test eventually adopted was to group the terms not distinguished by the clump classification, and look at these. Accordingly, for each term, a list of the clumps to which it belongs was prepared, and groups of terms were found which had all their clumps in common. These groups were quite small (2-6 terms) and could be studied easily. It turned out that some groups were ones of which a human classifier could have thought (e. g., words concerning suffix removal for machine translation came together) while others were quite justified by the documents concerned, but would never have been thought of a priori. For example, the group: "phrase marker, phoneme, Markov process, terminal language" was entirely justified by the... contents of the library. It is groups of the latter kind that represent a success for clump theory, for they function usefully in retrieval but in no way form part of the structure of thought... which the human classifier's work is likely to reflect." [1]

Still another application of the theory of clumps may be of use in the construction of thesauri (Sparck-Jones, 1962 [564]. Here the assumption is that rows of a correlation matrix can be formed for words giving other words which are synonymous with respect to meaning. The overlaps of the same word's occurrence in two or more rows can then be used to find clumps which are presumed to represent conceptual groupings.

Applications of clump theory to problems of mechanized documentation are also being investigated by Dale and Dale of the Linguistics Research Center, the University of Texas. [2] They have begun experimentation to derive clumps for the 90 clue words used by Borko and the 260 source-item computer abstracts used by both Maron and Borko. Preliminary results reported so far are principally limited to considerations of the associative networks between terms as derived from the structure of the clumps discovered by several clump definitions. Mention should also be made of the work of Meetham and Vaswani at the National Physical Laboratory, Teddington, England, looking toward the use of similar techniques for machine-generated index vocabularies, with preliminary emphasis on testing them against a "library" consisting of the propositions of Euclid's geometry. [3]

[1] Needham, 1963 [431], p. 285-286.

[2] Dale and Dale, an unpublished report dated February 1964, [147].

[3] National Science Foundation's CR&D report No. 11, [430], p. 137; and Meetham, 1963 [413].

5.3 Latent Class Analysis

Like the earlier work of Tanimoto, the latent class analysis approach of Baker (1962 [27])to problems of automatic information classification and retrieval is at least to date theoretical rather than experimental in nature, and so will be considered only briefly here. Baker claims that the latent class model developed in the field of the sociological sciences for the determination of latent classes among individuals responding "yes" or "no" to items in a questionnaire would have attractive features for application to information categorization and search, because the model is based upon response patterns that are analogous to the presence or absence of clue words or phrases in documents and because the analysis yields an ordering ratio that could serve a function similar to the relevance weightings suggested by Maron and Kuhns.

This ordering ratio is the probability that a given pattern of clue words will occur in a document properly belonging to a particular latent class. The probabilities of the same pattern being generated by a document properly belonging to other classes are also provided, giving an uncertainty which Baker thinks justifiable because a "document could generate a given pattern of key words, yet not belong to the same area of interest as the majority of documents possessing the same pattern of keywords". [1/] It should be noted, however, that the question of how to select appropriate clue words is begged [2/] and that no computer programs are as yet available for carrying out latent class analyses. [3/]

5.4 Examples of Other Proposed Classificatory Techniques

There are certain other document classificatory techniques that have been proposed and to some extent investigated experimentally. Trials of document clusterings based on co-citingness, co-citedness, or bibliographic coupling as compared with subject content groupings have, as noted above, been conducted both by Kessler at the M. I. T. Libraries and by Salton's group at Harvard.[4/] Consideration of Doyle's work on word co-occurrence statistics has been deliberately deferred to a later section which covers his general "association map" approach. Similarly, several other investigations will be discussed in terms of potentially related research such as linguistic data processing.

Two particular examples of other suggested classificatory techniques for document grouping or classification are somewhat unusual, however. These are the methods proposed by Te Nuyl and by Lefkovitz (1963 [353]). Cleverdon and Mills comment on Te Nuyl's method as follows:

[1/] Baker, 1962 [27], p. 518.

[2/] Ibid, p. 517. Note also that the footnote states: "A referee of this paper has properly cautioned that the effectiven ess of an information retrieval system may be due more to the appropriateness of the key words than the subsequent processing." See also Hillman, 1963 [272], p. 323: "Baker's theory, however, is based on interrelationships of key words, and thus constitutes an approach which is regarded with some suspicion by Farradane, who thinks that the real problem concerns the interrelationships of the concepts which key words denote."

[3/] Baker, 1962 [27], p. 516.

[4/] See Kessler, 1963 [320]; Lesk, 1963 [356, 357], and p. 30 of this report.

> "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.