Modifications to derivative indexing techniques that tend toward normalizations of terminology and word usage, and increasingly sophisticated proposals for machine use of syntactic, semantic, and contextual clues hold out the promise of transition to more truly "subject" indexing and to automatic assignment indexing systems.

4. AUTOMATIC ASSIGNMENT INDEXING TECHNIQUES

Answers to the question of whether indexing by machine is possible are actually dependent in part on how the question of whether what can be achieved by machine is or is not properly termed "indexing" is answered. If "indexing" is defined as being more than the mere extraction of words from titles, abstracts, or text, then automatic derivative indexing, even when augmented by various modifications, normalizations, and editings, does not provide affirmative evidence. In the case of concept-oriented definitions of indexing, the question becomes one of whether or not automatic assignment indexing is possible. Experimental evidence suggesting that it is will be presented in this section.

We should note first, however, that just as there are differences of opinion as to what "indexing" means so there are similar differences, with respect to whether or not it represents concepts rather than extracted words. There are also a number of conflicting definitions of what is meant by "indexing" in contradistinction to "classifying". For some, the latter difference is related to questions of the number of labels or surrogates assigned to a single item to represent its subject contents, ranging from the assignment of a single subject category in a classification scheme involving mutually exclusive classes to the assignment of a number of terms or descriptor each standing for one of a number of aspects of the subject. For our purposes, however, we shall regard both the case of indexing with a number of descriptors and that of classifying to a single category or subject heading as being within the province of automatic assignment indexing, reserving the term "automatic classification" for the case where the machine is used to establish the classification or categorization scheme itself.

Actual experiments in automatic assignment indexing by Borko, Borko and Bernick, Maron, Salton, Stevens and Urban, Swanson, and Williams will be discussed briefly below. These discussions are generally in chronological order with respect to first reporting of results, except that the Salton-Lesk-Storm work reflects a somewhat different principle of assignment from the methods using clue word approaches and it is therefore described after these others have been discussed. Some of the similarities and differences between the various methods are then indicated. A brief final subsection covers related assignment indexing proposals for which experimental data is not available or has not as yet been reported in the literature.

4.1 Swanson and Later Work at Thompson Ramo-Wooldridge

Research on fully automatic indexing as well as on full text searching and retrieval at the Ramo-Wooldridge Corporation has been reported as being under way at least as early as the spring of 1958. 1/ As described elsewhere in this report, experiments in search and retrieval based upon full natural language text had used as test items short articles in the field of nuclear physics. In additional experiments representing a preliminary "clue word" approach to possibilities for automatic indexing procedures, some of this same material was used.

1/ National Science Foundation's CR&D rept. no. 2, [430], p. 32.

In these additional experiments, 27 articles in the nuclear physics subject area were included in a corpus of 100 articles, the remainder covering a variety of topics. Frequency counts of word occurrences for the physics material were obtained and the 12 most frequent words that were judged to be discriminatory for the subject were selected. The hypothesis was then tested, that if any document pertained to nuclear physics it would contain at least two of these words. Retrieval was achieved for 25 of the 27 documents and the two "irrelevant" documents also retrieved did include information at least peripherally related to the subject. It was thus evident that the retrieval effectiveness of automatic recognition of nuclear physics subject material in the general collection was considerably greater than the average effectiveness of retrieving responses to the highly specific search questions in nuclear physics that had been used in the full text searching experiments (Swanson, 1961 [586]).

This second set of experiments provided a transition from the full text searching work, which if it can be considered indexing at all is obviously derivative indexing, to work in the application of an automatic assignment indexing method to 1,200 newspaper clippings (Swanson, 1962 [584], 1963 [580]). These were brief news items for which machine-readable texts in the form of punched paper tape were available. Thesaurus-groups of words likely to be associated with each of 20 to 24 subject headings were first compiled on the basis of human analysis of 1,000 or more representative items. These word groups were further screened so that no word appeared in more than one group and so that each word retained should be uniquely indicative of the particular subject category. In the machine assignment procedure, subsequently, if a word occurs that belongs to a particular thesaurus group, the corresponding subject heading is assigned to the item in which that word occurs.

Results achieved with this technique appear to be highly promising, at least for this type of material. Swanson reports as follows:

> "Approximately 1,200 brief news items were classified into 20 nonhierarchical subject categories, both by a human and a machine procedure. Each item was assigned on the average to about four categories. The results of the two processes were compared. With the human process as a standard, the machine missed only seven percent of the correct subject assignments and made a number of irrelevant assignments equal to about 17 percent of the total. Nearly 40 percent of the automatic subject assignments judged finally to be correct were missed by the human catalogers."[1/]

While this accomplishment is actually due to the extensive human effort to compiling, organizing, and pruning of the uniquely indivative word lists, it is pointed out that this intellectual effort and the programming tasks need to be done only "once and for all".[2/] It is further pointed out that garbles or misspellings in the input text do not appear to affect the procedure, there being enough redundancy in the messages so that even if one or two clue words are missed, others will be present.[3/]

1/ Swanson, 1962 [584], p. 468.

2/ Ibid, p. 469.

3/ Swanson, 1963 [580], p. 5.

Swanson and his TRW associates have further proposed extensions of the prespecified unique clue-word technique. For example, it is suggested that machine processes of comparing words of titles, subtitles and chapter headings to lists of possible subject heading can be extended in sophistication by machine lookups of synonym groups and of characteristic subject-word associations. 1/ Frequency weightings may be taken into account, and similar measures of association and subject-indicativeness may be developed for phrases as well as for individual words. 2/ In general, however, the apparent success of this clue-word technique in tests to date should be considered in the light of the special character of the items, their extreme brevity, and the high probability that the fact-word incidence involved in news reporting is not typical of less popular and less factually oriented materials. 3/

Continuing work along similar lines has been carried forward at Ramo-Wooldridge in the "Word Correlation and Automatic Indexing Program" sponsored by the Council on Library Resources (1959 [490] and [491]). Here, the objectives are to develop and apply clue-word techniques to material that is much more representative of the scientific and technical literature. The thesaurus-groups, now called "indexonym" groups, are made up of words and phrases selected by extensive human analysis as being significantly "useful-for-retrieval-purposes".

New items would be processed in a word and phrase lookup operation, with each word or phrase being initially assigned the identifier number codes of all groups to which it belongs. However, unless a particular group's number is repeated several times within the space of a few paragraphs, it is not used as the basis for the actual assignment of an index tag. Provision would be made for calling human attention to items having a number of words that are not deleted by processing against a "useless-for-retrieval purposes" list, but that are not found in any of "accepted" groups. It is suggested that in this way it should be possible to "ascribe measures of automatically recognizable 'newness' to technical articles". 4/

4.2 Maron's Automatic Indexing Experiments

By April of 1959, the reports of work at Thompson Ramo-Wooldridge on automatic indexing and related problems submitted for the Current Research and Development in Scientific Documentation series included reference to Maron and a "probabilistic model for the assignment of index tags", as well as to Swanson's continuing projects. 5/

1/ Swanson, 1962 [584], p. 469.

2/ Swanson, 1963 [580], pp. 1-2.

3/ See also Mooers, 1963 [424].

4/ Thompson Ramo Wooldridge, 1959 [491], p. 2A.

5/ National Science Foundation's CR&D report No. 5 [430], p. 34.

In addition to his work on probabilistic indexing with emphasis on relevance weightings for index tags manually assigned, Maron has actively explored automatic assignment indexing chniques. The approach is also probabilistic, with emphasis on the statistics of association between content-indicative clue words and subject headings manually assigned to sample documents. The experimental corpus consisted of a group of abstracts in the field of computer technology indexed to 32 subject categories designed for the purposes of these investigations.

Common words such as articles and prepositions were first excluded. Next, words occurring less than three times were purged and words such as "data" and "computer" were also rejected because they occur so frequently in this literature. Approximately 1,000 words remained after these purging operations. After sorting the source documents to their most appropriate subject categories, statistical frequencies were obtained for the co-occurrences of the candidate clue-words with the categories and the resulting listings were manually examined to determine which words peaked in a particular category. Eventually, 90 such words were selected.

The occurrence of one or more of the 90 clue-words in the text of new documents was then used to predict the subject category to which the new item should belong. 1/ Tests were run with two groups of documents, one consisting of the source items from which the statistical frequency and word list data had been obtained, and the second group consisting of 145 genuinely new items. For the latter group, twenty documents contained no clue words whatever and forty items had only one. For the remaining 85 items having two or more clue words, the results of the computer assignment program were predictions of the correct category in 44, or 51.8 percent, of the cases. 2/ Results using the source documents were significantly better, as expected, with 84.6 percent accuracy of category prediction for 247 items. Results were also related to the number of clue words that occurred in the test items, with a prediction accuracy of only 48.7 percent for items with a single clue word rising to 100 percent probability of correct assignment if six or more clue words occurred.

Trachtenberg (1963 [608]) has also considered a probabilistic approach to automatic indexing and categorization of documents, similar to that of Maron. He suggests the investigation of two information theoretic measures with reference to determination of which of various possible clue words are significantly discriminating with respect to the different categories. He further suggests experiments using 90 clue words and the corpus used by both Maron and Borko, but no actual results have as yet been reported.

4.3 Automatic Indexing Investigations of Borko and Bernick

At the System Development Corporation, the work of Borko (1960 [73]), and of Borko and Bernick (1962 [77], 1963 [78], 1964 [79]) in the area of automatic indexing has involved both automatic assignment indexing and automatic classification techniques. They have not only reported actual indexing results but have provided data for the intercomparison of their techniques with the experiments of Maron for the same source material.

1/ Note that the word itself is not necessarily used as an index tag or label, as is the case for derivative indexing using an inclusion list approach. This is an important distinction.

2/ Maron, 1961 [395], p. 257.

The original Borko approach was based on the principles of factor analysis as these had been developed for the analysis of multivariate date, especially in the field of psychology. Borko's first experiments were directed to a corpus consisting of 618 abstracts in the field of psychology, amounting to approximately 50,000 words of total text and 6,800 different words. These words were sorted by computer program into an order reflecting their respective frequencies of occurrence. For the approximately 200 words that occurred twenty or more times in this corpus, the investigator himself selected 90 words to serve as index (or, better, index-clue) terms. A matrix was then developed for the frequencies of co-occurrence of these words and the documents in which they appeared. From this, a 90 x 90 correlation matrix was computed as follows:

"To compute the correlation coefficient ... we used the following formula

$$r_{xy} = \frac{N\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{[N\Sigma x^2 - (\Sigma x)^2][N\Sigma y^2 - (\Sigma y)^2]}}$$

Where N is equal to the number of documents (618) and x and y are the terms being correlated." 1/

The term-correlation matrix was then factor analyzed and the first ten eigenvectors were selected as factors to be rotated and interpreted. Borko emphasizes that:

"The interpretation must be made by the investigator and is based upon his knowledge of the analytic procedures and the subject matter. There is, therefore, a degree of subjectivity in the names selected for each factor. These names may be regarded as hypotheses about the factor meaning." 2/

Following the derivation of these "classification categories" by means of the factor analysis technique, new items may be assigned to the categories on the basis of words occurring in their texts (abstracts) in accordance with the following procedural steps:

"1. Each document, in machine readable form, is analyzed by the computer. A list of the index terms and their frequencies of occurrence in each document is recorded.

"2. The category or categories containing the index term is assigned a value equal to the product of the number of occurrences of the word in the abstract and the normalized factor loading of the word in the category. If more than one index term appears in a category, the products are summed.

"3. After each index term has been considered, the category having the highest numerical value is selected." 3/

1/ Borko, 1961 [73], p. 283.

2/ Ibid, pp. 285-286.

3/ Borko and Bernick, 1962 [77], pp. 7-8.

The choice of 90 clue words in Borko's work with abstracts in the field of psychological literature was apparently dictated by a matrix size which would be convenient for computer manipulation. [1] However, it happened to coincide with the number of clue words used by Maron in his experiments. Advantage was taken of this coincidence to obtain comparative data on the performance of the two assignment-indexing techniques as applied to the same material. The 260 computer literature abstracts used by Maron as source documents were processed to derive a correlation matrix for Maron's 90 manually selected words, which was then factor analyzed. Several sets of factors were extracted, rotated, and the results studied, with a final selection of 21 categories.

Since these automatically derived categories did not coincide with Maron's original 32, it was necessary to analyze manually the total group of 405 abstracts (260 "source" and 145 "test" items) and assign them to the new categories, then to study the documents falling into each factor-analytically derived category to determine which of Maron's 90 clue words were category-indicative, and finally to substitute these words in the Bayesian equation used by Maron so as to predict which of these classification categories his probabilistic method should obtain.

The same two sets of 260 "source" and 145 "new" abstracts used by Maron were then submitted to the computer assignment program which compares the clue words of a new item with the numeric values of the predictor words for each factor category, then computes the score for each item in all categories, and assigns the category with the highest score to the item. For the source items, Borko and Bernick's results showed 63.4 percent correctly classified, by comparison with the 84.6 percent correctness score originally obtained for them in Maron's experiments. For the new items the factor analysis method scored 48.9 percent correct assignment by comparison with Maron's original 51.8 percent. [2] The later investigators therefore concede that the performance of Maron's technique was somewhat superior for the same items using the clue words originally selected by Maron.

Further experimentation was then carried out (Borko and Bernick, 1963 [78]) using word frequency data for the selection of a new set of 90 clue words and a classification scheme for 21 categories was again automatically derived. The 405 abstracts were again manually classified to these machine-derived categories by five subject-matter specialists and the two investigators. Comparative data were then obtained for both the Maron assignment formula and the modified classification system assignments in terms of agreement with the manual assignments.

For the source items, the percentage of machine assignments agreeing with those made by people was 62.7 when the Bayesian probability formula used by Maron was applied and 61.2 for the factor analysis score system. For the new items, the corresponding correct percentages were 57.9 and 55.9. Additional data compared the effects of using the original Maron words and the frequency-based word set (Borko's words) for the same probability formula assignment method. While there was an overlap of approximately 50 percent between Maron's words and Borko's words, the findings indicated that:

---

[1] Now increased to 150 x 150.

[2] Borko and Bernick, 1962 [72], pp. 9-10.

"... The index words selected by Maron are decidedly specific to the documents from which they were derived and are of less generality than the frequency based terms. The Bayesian formula coupled with the Maron words correctly predicted the classification of 79.6% of the documents in Group I ['source items'] but only 45.5% of the documents in Group II ['test items']. The coupling of the Bayesian formula with the Borko words resulted in a slight decrease in the percentage of Group I documents whose classification was correctly predicted (62.7%) but increased the percentage of correct prediction for Group II documents to 58.0%." 1/

Other findings from the later experiments indicated that despite the differences in the two word-sets, the factor categories derived from them were very similar. It was also found that, at least for the source items (Group I), the two machine techniques and the manual process classified 56.1 percent of the items into the same categories. It should be noted, however, that in the case of the automatic assignment methods: "Eleven documents contained no clue words and could not be automatically classified by either system." 2/

4.4 Williams' Discriminant Analysis Method

The work of Williams in automatic assignment indexing, reported in the fall of 1963 [642], has also involved tests on abstracts of the computer literature, directly comparable to but not necessarily identical with those used by Maron and by Borko and Bernick. This work at IBM's Federal Systems Division, Bethesda is based in part on earlier work by Meadow which involved computer studies of matching functions for document word lists and category word lists for test items drawn from such fields as psychology, law, computer abstracts, and news items. 3/ What has subsequently been developed is termed a "discriminant" method which begins with hierarchical classification structure of pre-established subject categories and with a small set of sample documents previously indexed by people into these categories. Frequency counts of words in each of the sample documents lead to computations, for each category, of the theoretically probable frequencies of its most statistically significant words. For new items, observed word frequencies are compared with the theoretical word-category associations and a relevance value is computed for the item in terms of each category.

The corpus selected for experimentation consisted of 400 items from "Computer Abstracts on Cards". 4/ These had previously been indexed using a classification structure of 15 major categories, each of which is divided in turn into 10 subcategories. The experimental sample, however, was so selected as to provide exactly 15 "source" items and 5 "new" items for each of 5 subdivisions of 4 of these major categories.

1/ Borko and Bernick, 1963 [78], p. 23.

2/ Ibid, p. 11.

3/ Williams, 1963 [642], cites H. R. Meadow, "Statistical Analysis and Classification of Documents", IRAD Task No. 0353, FSD IBM, Rockville, Maryland, 1962, but this is apparently a company-confidential document, containing proprietary information. Meadow gave an informal report on her work at the Computing Center seminars, University of Maryland, in March of 1963.

4/ Available on a subscription basis from Cambridge Communications Corporation, Cambridge, Mass.

Discriminant coefficients were then computed at both the major and minor levels for all words occurring in the sample items falling into one of the 20 groups in accordance with the formula:

"The discriminant coefficient is:

$$\lambda_i = \sum_j^n \frac{(P_{ij} - \bar{P}_{ij})^2}{\bar{P}_{ij}}$$

Where:

$$P_{ij} = f_{ij} / \sum_i^m$$

The relative frequency of the ith word in the jth category.

and

$$\bar{P}_{ij} = \frac{1}{n} \sum_j^n P_{ij}$$

The mean relative frequency per category of the ith word. 1/

These coefficients are used both to set up threshold values to determine which words should be used in the assignment formulas and to assign weighting factors to the words themselves.

The results of the experiments to date are based on 83 items from the "reference set" which were not used as source items. For 63 items, 78 percent were correctly classified at the level of a single major category (e.g., "Programming", "Hardware Design") and also correctly classified at a single subcategory level, (e.g., "Programming Languages", "Semiconductor Devices"). The 20 remaining items were classified to one major category with an accuracy of 95 percent and to two minor level subdivisions with accuracies of 60 percent and 75 percent. Additional investigations were made on the effects of using a discrimination threshold to eliminate insignificant words from consideration and on the use of weighting factors in the assignment calculations.

4.5 SADSACT

Stevens and Urban at the National Bureau of Standards (1963 [569, 570]) have also explored an automatic indexing technique that uses, as in the experiments of Williams, a teaching sample or reference set of previously indexed items to form patterns of word and index-term assignment associations. However, there are much less formal requirements for computing correlation coefficients and no consideration is required of either

---

1/ Williams 1963 [642], p. 163.

the theoretical probabilities of word occurrence by category or of discrimination coefficients and thresholds. Instead, the technique involves ad hoc statistical associations between the words occurring in the title and in the abstract of a sample item and the descriptors previously assigned to that item. A master selection-word vocabulary is thus built up where each word is listed in terms of the frequencies of its co-occurrence with each of the descriptors with which it has co-occurred, regardless of whether or not such prior associations are either revelant or significant. No attempt has as yet been made to "purge" the resulting association lists. Instead, reliance is placed on the patterns of multiple word usage and of redundancy of words used in titles and cited titles of new items to minimize the effects of irrelevant or accidental prior word-descriptor associations and to enhance the significant ones.

The SADSACT method (for "Self Assigned Descriptors from Self and Cited Titles") proceeds with the assumption, which it shares with the arguments for citation indexing previously discussed, that the literature references cited by an author are indicative of the subject content or contents of his paper. 1/ For the automatic indexing of new items, their titles and the titles of up to ten bibliographic references cited are keystroked, converted to punched cards, and fed to the computer. This input material is run against the master vocabulary to obtain for each input word which matches a vocabulary word a "descriptor-selection score" for each of the descriptors previously associated with that word. These scores are summed up for all words and at an appropriate cutting level those descriptors having the highest scores are assigned to the new item.

Preliminary results based on the titles and cited titles of items that were "source items" in the sense that their titles and abstracts had been used in the teaching sample were reported at the NATO Advanced Study Institute on Automatic Document Analysis held in Venice in July, 1963. For 30 items drawn from such subject fields as computer technology, information selection and retrieval, mathematical logic, pattern recognition, and operations research, all of which had previously been indexed by ASTIA personnel in 1960, the machine assigned 64.8 percent of the descriptors previously assigned. Subsequent tests on genuinely new items, however, resulted in a drop to only 48.2 percent "hit" accuracy.

These "new" item results were also evaluated by having several representative users of the collection analyze the test items and assign descriptors to them from a list of the descriptors available to the machine. The extent to which the descriptors assigned by machine were also independently chosen by one or more of these indexers was then checked. In general, the fewer descriptors assigned by the machine, the better was the human agreement, ranging from 47.4 percent overall in the case where the machine had assigned twelve descriptors to each item to 76% agreement where the machine assigned only one. In particular, for ten items which were analyzed by five different indexers, the chances that one or more would also select the machine's first choice (highest scoring) descriptor averaged 90 percent.

4.6 Assignment Indexing from Citation Data

Certain phases in the program of investigation of information selection and retrieval problems at the Harvard Computation Laboratory have been mentioned previously. The work of Storm and of Lesk and Storm on the use of first-noun-occurrences as selection clues for both automatic indexing and abstracting was discussed in connection with techniques for improved derivative indexing. The studies on citation indexing have included, as noted, experiments to assign indexing terms to a new document by finding the indexing

1/ If necessary or desirable, however, abstracts or portions of text can be used in addition to or in lieu of the cited titles.

terms previously assigned to the five most "related" documents, where "relatedness" is a function of the similarity in citation patterns as between the new document and items already in the collection. The results of such index term assignments are reported as identical to those made by human judgment approximately 50 percent of the time. 1/

More specifically, in an experiment using documents drawn from a small collection in the fields of mathematical linguistics and machine translation, a new item was compared in terms of its citation data with the citation similarity data previously determined for earlier documents, and the set of five related documents was selected using the magnitude of the row similarity coefficients obtained from links of length one and two. All index terms occurring at least twice in the set of terms assigned to these related items were then assigned to tne new items. For the ten "typical" new item cases, for which comparative data are shown, the citation data assignment method correctly assigned, on average, 47.6 percent of the terms assigned manually to the same items. 2/

A slightly more sophisticated indexing term assignment formula, described by Lesk, was applied to additional test cases, but "failed to raise accuracy above fifty percent". 3/ For five typical new cases, the improved method correctly assigned 11 of the 20 terms manually assigned to these items, or an average accuracy of 55.5 percent. 4/

4.7 Similarities and Distinctions among Assignment Indexing Experiments.

In Table 2 some of the key points of the various automatic assignment indexing experiments we have discussed above are summarized. Certain similarities, distinctions, and differences are to be noted. Borko and Bernick use the same corpus as did Maron and also re-apply Maron's formula to a different clue-word set for the same material. Williams uses material similar to the Maron-Borko computer corpus. The SADSACT tests also use some items that might be included in the Maron-Borko and Williams corpora. The Swanson experiments with newspaper clippings represent a quite different class of material consisting of brief, terse, factual messages.

1/ Lesk, 1963 [357], p. V-8.

2/ Salton, 1962 [520], p. III-41, Table 9.

3/ Lesk 1963 [357], p. V-7.

4/ Ibid, p. V-8, Table 3.

Table 2. Summary of Automatic Assignment Indexing Test Evaluations

Investigator	Principles and Methods	Materials Used	Tests	Remarks
Maron	Statistical probabilities of association between clue words and pre-established subject categories. Source items manually indexed to 32 categories. A subclass of words occurring in the corpus selected as clue words, and statistical correlations obtained for 90 such words with categories assigned. Correlation data and Bayesian probabilities used to assign categories to new items.	Corpus of 405 items selected from computer abstracts, PGEC, 1959. Full text, 20,000 words of which 3,263 were different words.	For 260 source items, 12 did not contain any clue words, 247 were indexed, 1 contained an error preventing processing. For the 247 source items indexed, probability of top-ranked category being correct = 84.6%. For 145 new items, 20 not indexed because they contained no clue words. In 85 cases where at least 2 clue words occurred, probability of correct category assignment = 51.8%.	Considerable manual inspection and judgment involved in the selection of clue words. Some new items cannot be processed, because they contain no clue words.
Borko	Factor analysis to determine distinctive grouping of clue words. Word frequency counts made, 90 of the 2.0 most frequent non-common words manually selected. Correlation matrix computed, factors rotated and interpreted.	Psychological abstracts. 618 abstracts, 50,000 text words; 6,800 different words.	Factors selected were judged to be compatible with but not identical to subject classification terms used for these items by the American Psychological Association.	Some new items cannot be processed, because they contain no clue words.

Table 2 (cont.)

Investigator	Principles and Methods	Materials Used	Tests	Remarks
Borko and Bernick	Factor analysis to determine distinctive groupings of clue words. Maron's 90 clue words used for word-word correlation and factor analysis. 21 factors developed, and items manually re-indexed to these categories.	Same corpus as Maron, 405 computer abstracts, of which 260 used to establish factors, 145 as new items.	Detailed comparison with Maron's technique. For the source items, 63.4% were correctly classified. For the new items, 46.5% correctly indexed, and 48.9% were correct for those items in which 2 or more clue words occurred.	Some items cannot be processed because they contain no clue words.
Swanson	Text word lookup against clue word lists, constructed by careful analysis of sample items to be exclusively indicative of a particular subject heading. Machine assigns a subject heading to an item if any word on its list occurs in that item.	Brief news dispatches available on teletype tape, wide diversity of topics. From study of several 1,000 items, 24 subject headings established and word lists selected, averaging approximately one hundred per category. 775 new items then tested.	Machine assignments compared to manual subject indexing. For a first batch of 500 items, 569 assignments of correct headings, 119 assignments of irrelevant headings, and 32 correct headings missed. The clue word thesaurus was then revised. For 275 additional test items, results showed 282 correct assignments, 29 irrelevant assignments, 1 missed. For total, averages of 17% irrelevant assignments, 3% missed. For 200 items, machine and manual assignments were compared with respect to 5 of the subject categories, with the following results: Man / Machine Irrelevant 4 / 25 missed 46 / 4 correct 75 / 116	

Table 2 (cont.)

Investigator	Principles and Methods	Materials Used	Tests	Remarks
Stevens and Urban	Teaching sample for machine compilation of co-occurrence data for words in titles and abstracts with descriptors assigned to these items. Words in titles and cited titles of new items then run against master list of previous word-descriptor association to derive descriptor-selection scores, highest scoring descriptors (e.g., up to 12) assigned. Associations derived for 1,600 words co-occurring with any of 70 descriptors previously assigned.	Two teaching samples, approximately 100 items each with 70% overlap, drawn from items indexed by ASTIA. For new items titles and up to 10 cited titles.	For 59 test items, assignments of descriptors that had occurred for at least 3% of the sample items agreed with ASTIA assignments 58.1%. However, for all descriptors assigned by ASTIA, many not available to machine, overall machine accuracy = 40.1%. For 20 items, independently evaluated by several typical users, the chances that one or more people would agree with the machine assignments ranged from 47.1% when 12 descriptors were assigned to 75.0% average agreement with the machine's first choice.	All test items could be processed and up to 12 different descriptors assigned to each, but some descriptors used in manual indexing of these items are not available to the machine.
Williams	Discriminant analysis. Sample items previously indexed to a 2-level classification system were subjected to word frequency counts and the theoretical frequencies of the most significant words in each category were compiled. For new items, observed word frequencies' compared with theoretical frequencies for each category, highest scoring assigned.	Items from "Computer Abstracts on Cards" indexed to 15 major categories each divided into 10 minor categories. 300 abstracts selected to provide equal distribution to 20 sub-categories, 5 each in 4 major categories. Additional items for test similarly selected.	For 63 new items assigned by machine to 1 major and 1 minor category, 78% correct at major level, 64% correct at minor level. For 20 items classified to 1 major and 2 minor categories, 95% correct at major level, 60% and 75% correct at the minor level.	

None of the experiments has so far encompassed testing of anything but very small test item samples and the dangers of extrapolating from so small and so specialized bodies of data should be clearly recognized. Mooers identifies these dangers in terms of

"The Silent Postulate:

That (real people) (real documents) (real jobs to do) can somehow

be eliminated from the experimental study, and that (role-playing people) (substitute documents) (imaginery jobs)

can be substituted and still give valid experimental results." 1/

In most of the experiments in automatic indexing conducted to date, indexing and classification schedules have been especially designed, or evaluations made, specifically for the purposes of these tests. Williams, however, stresses the point that the material used in his experiments had been "classified by professional indexers for the purposes of actual retrieval." 2/ A similar claim can be made for SADSACT, as noted by Mooers. 3/ Swanson's news item work also obviously relates to real items and implies a real job to be done, but is directed, as noted, to a class of material not generally comparable to that found in documentation operations on scientific and technical literature.

In contrast with the treatment of each document as a self-contained entity without reference to any other documents, as is the case for derivative indexing, all of the automatic assignment indexing experiments, by virtue of the fact that they are assignment techniques, do to some extent embody the effects of a consensus of a particular collection, or a consensus of prior indexing, or a consensus of human subject content analysis applied to sample documents, or some combination of these effects. The SADSACT method, in addition, wherever cited titles are available for new items, takes advantage of terminology other than the author's own as a source of clue words. Other proposed methods of assignment indexing, such as the use by Salton, Lesk, and Storm of citation-pattern similarity data, would carry the latter principle even further.

1/ Mooers, 1963 [424], p. 5.

2/ Williams, 1963 [642], p. 162.

3/ Ibid, p. 5.

4.8 Other Assignment Indexing Proposals

A few additional automatic assignment indexing proposals are under development. Examples for which experimental data is not as yet generally available include, for example, work at EURATOM, some preliminary experiments at Chemical Abstracts Service, work at General Electric, Bethesda, the proposed "Multilindex" system of Information Systems, Inc., investigations by Slamecka and Zunde, and a special purpose development project at Goodyear Aerospace.

Meyer-Uhlenried and Lustig report for the EURATOM developments as follows:

> "... Procedures are being developed which allow based upon given keyword lists first for abstracts: (a) to assign significant keywords and (b) based upon hierarchically organized keyword lists, to assign the documents in question to specific subject fields.
>
> "Experiments were made at first on narrow fields with so-called micro-thesauri, they showed encouraging results when automatic and manual assignment were compared. Positive results depend of course on the quality of the abstracts and the significance of the words employed in them. It remains to see how far this favorable prognosis is confirmed by keyword collections of more complex contents." 1/

Friedman and Dyson (1961 [203]) have reported on manual experiments designed to relate words occurring in a sample of abstracts from a particular section of Chemical Abstracts to the title or heading for that section. Significant words in these abstracts were counted and the number of occurrences as well as the number of different abstracts in which they appeared were determined, with a rank order listing as a result. It appeared, from inspection, that it should be feasible to develop, for each CA section, a relatively small vocabulary of words that would be descriptive, and indicative of, the subject matter contained in it. They conclude: "In our opinion, the results were significant, the small vocabulary of words did select a large percentage of the abstracts in the section it was based on." 2/

A project at Information Systems Operations, General Electric, on possibilities for automatic indexing and abstracting of text has been reported in the November 1962 issue of Current Research and Development.3/The META project (Methods of Extracting Text Automatically) is said to be concerned with the use of statistical, linguistic, and semantic criteria for analysis and selection of significant words and significant sentences from text. Computer programs are being developed in modular fashion for the GE-225 computer.

1/ Meyer-Uhlenried and Lustig, 1963 [417], p. 229.

2/ Friedman and Dyson, 1961 [203], p. 10.

3/ National Science Foundation's CR&D report, No. 11 [430], p. 97.

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.