9. CONCLUSION: APPRAISAL OF THE STATE OF THE ART IN AUTOMATIC INDEXING

Notwithstanding the difficulties of evaluation we have discussed, we shall herewith attempt to evaluate the present state of the art in automatic indexing techniques, using such available criteria as seem most appropriate. First, we suggest that all of out initial questions except possibly the last, can today be answered affirmatively. "Is indexing by machine possible at all?" To this we can answer an unequivocal "yes" in view of the many examples of KWIC type indexes extant and in practical use. Secondly, "Is what can be done by machine properly termed 'abstracting', 'indexing', or 'classifying'?" If, by definition, word indexing of any kind is not "properly termed... indexing", then, as we have seen, automatic derivative indexing, such as KWIC, or the selection of words to serve as index tags based upon the frequencies of their occurrence in text, is not so either.

The fundamental Luhn concept for indexing based on word frequencies is, as we have seen, straightforward: namely that, after disregarding the most frequent "common words", especially those that are syntactic-function words -- articles, conjunctions, prepositions, and the like, together with those words that occur infrequently in a given text, the remaining high frequency words should give a reasonable indication of what the author was writing "about". Critiques of the Luhn position have been made on several-fold grounds:

(1) Information-theoretic - that, in fact, the most information is conveyed by the least frequent words.
(2) Absolute vs. relative frequencies of usage within specialized fields.
(3) Modifications of semantic purport by contextual and syntactic associations.
(4) Problems of synonymity and, conversely, of orthographically identical words. 1/
(5) Multi-aspect points of interest, and future need of access to material the author himself did not emphasize.

The last point raises again the criticisms that have been made against derivative, extractive or "word" indexing of all types. To repeat, although such procedures may index "as the author himself indexed best -- in his own language", the significant points are (1) there may be peripheral, minor, or unrecognized aspects of his topic and incidental information disclosed, of future interest to others, which the author himself is in no special position to recognize, and (2) notwithstanding the "author's own terminology" being current usage rather than the "fossilized" vocabulary of any previously established classification or indexing scheme, this very "currency" changes from field to field and, quite literally, from day to day. Nevertheless, it should be re-emphasized that the validity of these criticisms is not limited to automatic derivative indexing as such, but rather is applicable against any indexing system whatsoever, manual or machine, which is so strictly limited to author-terminology, author-emphases, and the consideration of the document at hand as a self-contained entity, without regard to other documents in a collection, in a particular field, and without respect to specific user needs. By contrast to this type of limitation, more promising approaches should stress both similarities and differences between a new document and previously received documents, between documents "belonging" to some definable category, or not, and even, as responsive to a particular user's profile-of-interest, or not.

1/ See Baxendale, 1962 [42], pp. 67-68: "... resolution of orthographic ambiguities is a non-trivial and over-riding prerequisite for the computer processing of text...", p. 67.

Derivative indexing, whether by man or machine, is thus subject to many disadvantages. First and foremost, it is constrained by a particular individual's personal manner of expression of concepts in language. This limitation is controlled only by his presumptive desire to communicate with some particular (more or less general, or more or less specialized) audience. His choices of natural language expressions, however, will be conditioned by at least some of the following factors:

(1) The range and precision of his personal mastery of both general and specialized vocabularies for a given time, place, and specialized field of discourse.
(2) His personal expectations as to the probable reactions (in the sense of effective communication) of his intended audience to the expressions that he does choose, involving all of the problems of different usages of technical terminology from field to field, from formal to informal presentations, from scholarly reviews to progress reports heavy in current "technese" and "fashionable words".
(3) His habits of thought and his training in his field.
(4) His awareness of more than one possible audience and of more than one point or topic of potential interest to his readers.

Secondly, indexing by the author's own words is remarkably sensitive to a particular period of time, so that the terminology becomes rapidly outdated and often seriously misleading in its connotations. Thirdly, the user has no advance knowledge of the terminology that has been used in all the varied texts of a collection and he must therefore be able to predict a wide variety of possible ways of expressing ideas in words, phrases, and even by implication. Fourthly, for collections indexed on a word-derivative basis, there is little or no possibility for generic searching. 1/ Finally, there is the more general question, applicable to both derivative and assignment indexing, of how well, ever, can a condensed representation serve the purposes of specific subject content recapture? In the strict sense, only by the elimination of truly redundant information. But even this is a relative matter. What is redundant for an author may not be so for several different potential users of the reports or papers that this author writes. What is redundant for one user is not necessarily so for others.

The further problem for machine techniques is therefore: how selection rules can be provided that will replicate a given human pattern of selectivity, or, alternatively, how selection rules can be established and defined that will produce an equivalent and comparable result - that is, one which typical users would agree is as pertinent to their query-answer relevance decisions as any available alternative.

Certainly the problem of appropriate selection is at the heart of the matter. This is a crucial question, even if we sort out and can specify the different uses, for a particular collection, a particular clientele, at a particular time, that automatically generated condensed document representations may have. Wyllys, in appraising automatic abstracting efforts, considers that the goal should be to provide extracts which will serve a search-tool function -- that is, they will furnish the searcher with enough information about the document content so that he may decide whether it is probably pertinent to his then interests or not and hence decide whether or not to read the document in full. By contrast, he says of the "content-revelatory function" that an abstract should: "furnish the reader with enough information about the related document so that in most cases he will not need to read it itself." 2/

1/ See for example, Doyle, 1963 [162], with respect to lack of capacity for generic searching as one of the major disadvantages of natural text search systems.

2/ Wyllys, 1963 [653], p. 6.

Let us recall the objections to the use of the terms "auto-encoding" (or "auto-indexing" or "auto-abstracting") because of the possible connotation of self-encoding, etc.. 1/ This is an objection based upon avoiding ambiguous or misleading terminology, but it also points to an objection as to the principle involved--that is, of treating the document itself, in its own right, as a self-sufficient, self-contained, universe of discourse, and of assuming that some type of summation-condensation over a number of different and individually-derived representations of the separate documents in a collection can provide an effective selection-retrieval guidance system to the contents of various specific documents in that collection. Even when the actual operations are to be abetted by synonym reduction and normalization procedures (whether at the indexing or search negotiation stage, or both), there is a significant difference between this endogenous hypothesis and its exogenous alternative: that the basis for automatic indexing be the consensus of the collection, or of a sample of the collection, or of prior indexing.

Assignment indexing, especially in the sense that concept-indexing is the goal, may be subjectively preferable to derivative indexing not only because it involves exogenous emphases but because it tends to delimit, centralize, and standardize the access points available to the user in his search-retrieval operations. However, in terms of the human indexing situation, it involves all the traditional difficulties of indexing - which in turn invoke the problems of evaluating indexing systems:

> "Justification for any indexing technique must ultimately be based on successful retrieval. Success can only be evaluated in terms of a closed system; that is, a system wherein sufficient knowledge is available of the entire contents of the materials, so that an evaluation can be made of various techniques as to their retrieval effectiveness. The various systems ... cannot really be weighed except on the basis of a test comparing one against the other. This has not been done in any place." 2/

Nevertheless, there are a variety of reasons for accepting even the relatively crude derivative indexing products as practical tools today, for seeking machine-usable rules for the improvement of these products, and for continuing research efforts in automatic assignment indexing and automatic classification. There are, first and foremost, the cases where conventional indexes are inadequate or non-existent. Thus Wyllys claims:

> "It is well-known that the current methods of producing, through human efforts, condensed representations of documents are already hopelessly inadequate to cope with the present volume of scientific and technical literature. Many papers are never indexed or abstracted at all, and even in the cases of those that are indexed or abstracted, the indexes and abstracts do not become available until six months to two years after the publication of the paper." 3/

Again, with respect to automatic derivative indexing, especially KWIC indexes based on titles alone, there can be no question as to the evaluation criterion of timeliness. The success of this aspect is widely acknowledged by users, systems planners, and interested observers. On the other hand, there is very little reported evidence available on which

---

1/ See p. 3 of this report.

2/ Black, 1963 [64], p. 16.

3/ Wyllys, 1961 [650], p. 6.

any objective measure of comparative cost-benefit ratios may be obtained. Black reports, but without supporting data, that:

> "It has been estimated that the efficiency of KWIC indexing is about 76 per cent compared with about 82 per cent for conventional indexing or classification." 1/

White and Walsh report that:

> "From the limited experiment on methods of indexing the 1962 issues of the Abstracts of Computer Literature, the permuted title indexing retrieved only 52 percent of the information. This low percentage may be attributed to the changing and not yet uniformly standardized terminology existing in computer technology." 2/

KWIC indexes, because of their very currency, are fulfilling significant maintaining-awareness needs today. Improved titling practice, enforced by editorial rigor or contractual requirements or both, can improve their usefulness. They fill gaps in the bench scientist's or engineer's ability to know about what might be of interest to him, either because the material is not otherwise covered in normal secondary publication (e.g., conferences and proceedings of symposia, internal technical reports not produced on Government contracts and therefore not announced and indexed by the cognizant agencies, and the like) or because the sheer bulk of the product of indexing-abstracting services in his field prevents his effective use of these services unless more specific access points are provided. The claim that "something is better than nothing" is not without merit, 3/ even with all the problems of non-resolution of synonymity, homography, topical scatter, long blocks of entries under the sorting term, the even more significant disadvantages of author-bias towards his principle topic, the author's choice both of emphasis and terminology, and the like. Williams, considering word-with-context indexes, whether limited to title only or to titles with readily available augmentation, makes the following comments:

> "Limitations and other troublesome features of the method have been obvious, but perhaps over obvious, in the light of its growing acceptance and of the basic validity of permitting a document to speak for itself, even in a much abstracted recapitulation. Wherever there are large and growing problems in maintaining publication schedules for established subject indexes, or wherever pressing needs develop for more frequent indexes, for rapid, low-cost cumulation, or for indexes in areas where suitable indexing services are wanting, there no apology is needed for proposing that this method be considered and tried, as a precursor to 'better' indexing, if not as a substitute. Its use may be of interest also in less troubled circumstances, in its own right, and because of common elements involved in its production and the provision of other wanted products and functions (catalog records, current-awareness, lists, etc)." 4/

Returning to the question of whether automatic indexing is possible, it can be seen that, at least in the derivative indexing sense, it is not only possible but can be practically useful. To dismiss the evidence of automatic derivative indexing operations that are in production today by rigorous definition of what indexing is in effect anticipates both our

---

1/ Black, 1962 [65], p. 318.

2/ White and Walsh, 1963 [639], p. 346.

3/ See Veilleux, 1962 [624], p. 81: "Accepting the premise that partial control of information satisfies more consumers than absence of control, perfection was traded for currency."

4/ T. M. Williams, private communication, dated January 4, 1962.

third and fourth questions: whether machine-generated indexes are as good or better than the products of human operations and of how we can measure and appraise the adequacy of any indexing system whatever. Here are encountered the "core" problems of meaning in communication, of information loss in any reductive transformation of actual messages or documents, of relevance of particular messages to particular queries and to particular human needs, of judgments of relevance.

Because of these underlying yet overriding questions, the state-of-the-art in the evaluation of indexing systems is in fact far more primitive than that of automatic indexing itself. An easy, and an early, solution is not likely. Therefore, today, in appraising machine potentials for assignment indexing we are faced with what is in effect a single criterion: namely, will a given group of human evaluators, whatever their standards and requirements, agree as much with the products of an automatic indexing procedure, otherwise competitive on a cost-benefit ratio with human indexing of the same material, as they do amongst themselves?

Within the limits of small, specially selected samples of document or message collections, it is possible to demonstrate that:

(1) Replication of the products of at least some existing systems, within the consistency levels observed for these systems, can be achieved.
(2) Retrieval effectiveness with respect to relevant items indexed by automatic assignment procedures can be at least as good as, and may be superior to, that obtained from run-of-the-mill manual indexing of the same items.
(3) Costs of indexing can be held at or below the costs of equivalent manual indexing, provided both that the input material required is already in machine-usable form, or can be held to an average of, say, 100 words or less, and that the clue-word lists, association factors, or probabilistic calculations can be accommodated within internal memory.
(4) Significant gains in time required to generate an index or to index or reindex a collection can be achieved.

Some degree of theoretical success in assignment indexing by machine can thus certainly be claimed. Moreover, many of the test results reported do clearly indicate a quality of indexing, for a given collection at a given level of specificity of indexing, at least comparable to that which is typically and routinely achieved by people in a practical indexing situation. No more should be asked of the automatic techniques unless better human indexing can be specified as being equally feasible, timely, and practical. Further, no more should be asked of automatic techniques in terms of the evaluation of their potentialities, than is now asked of the manually-prepared alternatives. 1/

Data with respect to comparison of the results of automatic assignment indexing techniques to either a priori or a posteriori human judgment have been mentioned previously in this report in terms of actual test results reported, and the most significant of these reported data are summarized in Table 2. 2/ Typically, however, these data reflect, in varying degrees, so small a sample of test cases, of user preferences, and/or of special purpose and interest, that no general extropolation is reasonable. Moreover, the general questions of the "core" problems of evaluation in general again rear their own ugly heads.

1/ Compare, for example, Kennedy, 1962 [311] and Needham, 1963 [433].

2/ See pp. 101-103 of this report.

Thus, Borko and Bernick point out:

> "Up to this point we have used human classification as our criterion for the accuracy of automatic document classification. Against this criterion we have been able to predict with approximately 55% accuracy, and no more. Is this because out techniques of automatic classification are not very good, or is it because our criterion of human classification is not very reliable? There is some evidence to indicate that the reliability of human indexers is not very high. The reliability of classifying technical reports needs investigating and, perhaps even more basically, the reasons for using human classification as a criterion at all." 1/

In general, the results of automatic index-term assignment procedures appear to run in the area of 45-75 percent agreement with prior human indexing, 2/ and this in turn is well within range of, and often superior to, estimates of human inter-indexer consistency based on actual observations and tests. There can be little or no doubt that the results of automatic assignment indexing experiments to date, (if extrapolation from the small and often highly specialized samples so far used in actual tests is in fact warranted 3/) do suggest that an indexing quality generally comparable to that achievable by run-of-the-mill manual operations, at comparable costs and with increased timeliness, can be achieved by machine.

The question which remains is simply that of practicality, today. Extrapolation from small samples is highly dangerous, as is well noted even by enthusiastis for machine techniques. The fact that for at least some systems, the limitations on number of clue words that can be handled (due in part to computational requirements, matrix manipulations, and the like) are such that, even in an experimental situation, certain "tests" are excluded from the result statistics, because the items contained an insufficient number of clues, is a serious indictment of reasonable extrapolations for these techniques today. Most tests so far reported have involved not only a highly specialized "sample" library or collection, but a severe limitation on the total number of "descriptors", subject headings, or classification categories to be assigned. Maron used 32, Borko 21, Williams 20, SADSACT 70, Swanson 24. How would any of these approaches fare, given several hundred, much less

---

1/ Borko and Bernick, 1963 [78], pp. 31-32.

2/ See Table 2.

3/ This is an important, perhaps crucial, caveat. See, for example, Goldwyn, 1963 [233], p. 321: "In the micro-experiments of many of those who would apply statistical techniques ... The document collection consists of 0-100 units. Results based on the manipulation, real or imagined, of such a collection can be valid for it, yet become shaky or even nonapplicable to larger collections"; Perry 1958 [471], p. 415: "A degree of selectivity quite acceptable for files of moderate size may prove quite inadequate in dealing with large files. This fact often makes it necessary to exert unusual care and considerable reserve in evaluating the results of small-scale tests and demonstrations which may tend to cause the mass effects of large files to be underestimated or overlooked completely"; Swanson, 1962 [586], p. 288: "The extent to which semantic characteristics of natural language are susceptible to being generalized from small sample data is deceptive."

several thousand, possible indexing or classificatory labels? 1/

The use of very brief short articles, or of abstracts, as the members of experimental corpora for investigations of automatic assignment indexing techniques presuming the processing of full text, either for indexing purposes or for subsequent "indexing-at-time-of search", is seriously misleading. First, it is not truly representative of discursive text, either in vocabulary-syntax, or stylistic variations involving synonymity, tropes, elisions, dangling referents, and inumerable other meaning-implications, not explicitly stated.

Secondly, as any author of a technical paper, for which he must provide an abstract, knows all too well, he must concentrate in the abstract on a telegraphic emphasis toward his principal topic and the points he wishes to make. He must omit most qualifying, specifying, and suggestive-of-other-leads-or-applications words and phrases, which he will in fact develop in the text itself. For this reason, even supposing that the author himself is unusually well-aware of the multiple points of access that many different potential users might desire, the required brevity of the abstract form almost necessarily demands terse, shorthand-type statements that can only increase the problems of "technese", of homography, and of single-subject representation.

Granted, in either manual or machine-serviceable systems today, the current-awareness scanning need is largely met by indexing based solely or primarily on title only, or title-plus-abstract. But is this good enough for search and retrieval? If and only if it is, then automatic indexing potentialities available today should be considered for both purposes.

Our final question as to whether automatic indexing can be accomplished by statistical means alone or must involve syntactic, semantic and pragmatic considerations is not entirely answerable. In terms of achieving comparable quality with many manually prepared indexes available today, statistical means alone do appear promising. But is the achievement of just this level (even if accompanied by significant gains in timeliness, coverage, and economy) really good enough? There are a number of serious investigators

---

1/ For example, Black predicts (1963) [64], p. 19) that for most systems an adequate vocabulary or thesaurus will comprise some twenty thousand terms. See also Arthur D. Little, Inc., 1963 [23], p. 65: "The enormous number of computations required increases very rapidly with the number of indexing terms. Existing computers, operating serially, do not appear to be capable of handling the problem economically for collections with 9000 or more terms even if the simplest associative techniques are employed"; Williams, 1963 [642], p. 162: "One of the practical problems... is in the inversion of large matrices. In certain methods the order of the matrix will equal the number of different word types in the population, which is usually in the thousands."

convinced that it is not, 1/ and for this reason, research efforts are being directed toward these other considerations.

On-going research and development work - whether in modified derivative indexing approaching a "concept-indexing" level; in automatic assignment indexing techniques as such; in automatic classification or categorization procedures, or in potentially related efforts directed toward automatic abstracting, automatic content analysis, and other aspects of linguistic data processing - is both reasonably extensive and quite promising. Most of the investigators who are seriously active in the field report their current objectives and recent accomplishments regularly to the National Science Foundation for publication in the series "Current Research and Development Efforts in Scientific Documentation." In the most recent issue, unfortunately current only as of November, 1962, there are not less than 25 reports of KWIC and similar title-permuted derivative indexing methods generated or proposed-to-be-generated by machine, there are several instances of investigations into various possibilities of modified derivative indexing to be accomplished by machine, and there are five to ten reports of active experimentation with various automatic assignment indexing schemes. These efforts and even more recently organized projects point in the hopeful direction that "KWIC indexes should be merely a sample of things to come". 2/

Assignment indexing techniques so far investigated can be, as we have seen, of two types which are quite distinct in terms of the principles involved. The first, which can be the more readily mechanized, involves the use of thesaurus-type lookup procedures covering the definable rules of "scope notes", "authority lists", or "see also" reference practice. The second type of assignment indexing, however, depends upon decision-making as to the propriety of assigning a particular indexing term to a particular document with reference to assignments to the collection as a whole (or a sample thereof). This latter type of assignment may be in terms of a priori categorizations of separable subsets of the collection.

Alternatively, the bases for the latter type assignment-indexing procedures may be derived from a posteriori determinations of the suitable subsets as in the factor analysis experiments of Borko, the latent class analysis approach of Baker, and the clustering-clumping approaches to automatic classification of Needham and others. It is to be noted in particular that Needham thinks an automatically generated categorization is preferable precisely because of lack of knowledge as to the exact attributes defining a class in

---

1/ See, for example, Climenson et al, 1962 [133], p. 178: "The statistical approach attempts to use no more than the occurrences of word spellings and their relative distances in the document environment ... [and] cannot provide the discrimination necessary for most indexing and abstracting applications"; Doyle, 1963 [162], p. 3: "Automatic indexing and abstracting, as currently conceived, do not require any sort of dictionary or other semantic reference, but only counting, comparing, and sorting-operations well known in numerical data processing. But success in applying such rules on a purely automatic basis can't help but be limited"; Borko, 1962 [75], p. 5: "Although difficult, identification [of different meanings carried by the same word, of the same meaning carried by different words] must be accomplished before the automatic categorization of document content can be truly effective. For the most part statistical methods, and even syntactic analysis, are inadequate for the job. A technique of textual analysis based upon the semantic properties of language is needed"; Grosch, 1959 [244], p. 20: "We need semantic methods ... that will look for the intersection of redundant descriptors, each of which is at least slightly erroneous."

2/ Doyle, 1962 [163], p. 381.

existing classification schemes. However, in the related field of pattern recognition Uhr and Vossler have shown promising results both for criterial feature analysis (a priori assumption as to attributes or properties governing membership in specified classes) and for randomly generated discrimination operators which, applied in a recursive manner, are increasingly adaptive to the detection of class-mambership (Uhr and Vossler, 1961 [615]).

One particular way of looking at the problems of automatic indexing results, in effect, in placing these problems within the broader field of pattern perception and pattern recognition. We suggest that this is in fact a particularly fruitful approach. Certainly there is a wide area of potential commonality, and many promising leads for further research in automatic categorization can be found in the general pattern recognition literature, especially in work on randomly generated operators and on the problems of determination of membership in classes. 1/ Conversely, automatic classification techniques originally conceived as applicable to the handling of documentary information have in fact been applied quite successfully to at least one case of groupings of physical objects on the bases of machine-detectable common properties.

The question of determination of membership-in-classes is basic to the problems of automatic classification and categorization. Thus the techniques for discriminating the statistically significant associations between "properties" of objects or items that are to be grouped into classes or categories, even when such "properties" are not known in advance and have no a priori identification, point to an increasing and promising convergence of research in pattern recognition, propaganda analysis and psycholinguistics, mathematics and statistics, studies of linear threshold devices, and the like, as well as in the linguistic data processing field as such.

It is true that such synthesized "classes" may have no convenient "names" or linguistic interpretations which make much sense to the individual human searcher or user. Nevertheless, what is suggested is that a radical departure from conventional habits of literature search and retrieval may be desirable from the standpoint of effective use of machine potentialities. This might mean that, ab initio, the customer would pose to the system a search query request not couched in his notion of words or terms actually used in the system, but either (a) an outline or statement of his own research proposal and plan of attack or (b) an indication of one or several items that he has already decided are pertinent to his interests, with a request for "more like these".

An equally radical departure from conventional present habits and thinking is already implicit in Needham's suggestion of an automatically derived classification system and manual assignments thereto. 2/ It would attack present-day machine capacity and processing time limitations such that property and class or category associations must be held to something less than 1,000 x 1,000, unless prohibitive processing costs are to be incurred. This approach would assume a one-time large-scale building of vocabulary and term or category associations and derivation of assignment algorithms, and the printing out of the results in multiple copies for use by low-level clerical personnel carrying out, indeed, "machine-like" indexing.

A final promising approach to the future prospects for fully automatic indexing and categorization is the perseverance in research and development efforts in advance of the

---

1/ See, for example, Sebesyten, 1961 [539], 1962 [538].

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

advent of versatile character readers and inexpensive, very large capacity, rapid direct access memories. These efforts will include not only further systematic exploration of syntactic, semantic and pragmatic considerations in linguistic data processing, but also further attacks on the problems of language and meaning themselves. Thus, we may conclude with Maron that: "automatic indexing represents the opening wedge in a general attack at not only the problems of identification search and retrieval, but also the problem of automatically transforming information on the basis of its content." 1/

If we are to attempt to solve this problem, as indeed we should, must we not look forward to the possibilities of rapid up-dating, thesaurus growth and revision, and quick and economical re-indexings of entire collections that only machine-processing capabilities can promise today?

ACKNOWLEDGEMENTS

The contributions of Miss Josephine L. Walkowicz and her staff in the preparation and checking of items for the bibliography, and of Mrs. Betty J. Anderson, Mrs. Helen B. Grantham, and Mrs. Anna K. Smilow in the typing and editing of the manuscript are gratefully acknowledged. The courtesy of Miss Thyllis Williams, Mr. Joseph Becker, Mr. Herbert Ohlman, and the late Hans Peter Luhn in making available unpublished materials is also gratefully acknowledged.

1/ Maron, 1961, [395], p. 240. See also Salton, 1962 [518], p. 234 and Borko and Bernick, 1962 [77], p. 3