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

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

7. PROBLEMS OF EVALUATION

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

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

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

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

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

> "If the answer turns out to be 'no', we might reasonably conclude that the only reliable and effective kind of human indexing is that which is already machine-like in nature." [1]

With a few noteworthy exceptions, there has been very little serious investigation of these problems and there is very little comparative data.

O'Connor has been making a series of studies, with considerable emphasis upon how one might measure the products of machine indexing and how one might derive machine rules for automatic index ing from systematic review of documents indexed by people. Cleverdon and his associates at the ASLIB Cranfield project have extensively tested several different indexing procedures. Painter, MacMillan and Welt, Slamecka and Zunde, and others report findings on intra-indexer, and inter-indexer consistency -- unfortunately, on the basis of quite small samples. Various alternate approaches to the evaluation of automatic indexing results have been considered by Borko, Doyle, Swanson, Savage, Giuliano, and others. In addition, some data bearing on these questions have been reported in connection with analyses of selective dissemination (SDI) systems. Some data from other sources, such as studies of user preferences with respect to various reference and search tools, is also pertinent.

The most generally accepted criterion for appraising the effectiveness o f indexing is that of retrieval effectiveness. But, in general, this is merely the substitution of one intangible for another, entailing a string of as yet unanswerable or at least unresolved questions.[2] Retrieval of what, for whom, and when? How can effectiveness be measured except by the elusive question of relevance judgments? How can human judgments of relevance and value be measured and quantified?

We shall try to distinguish here, insofar as possible, between the core problems that make the evaluation of indexing as such an extremely difficult task, the available data on human indexer reliability, and the possible advantages and disadvantages of automatic indexing techniques.

[1] Montgomery and Swanson, 1962 [421], p. 366.

[2] Compare Swanson, 1960 [582], pp. 2-3: "The performance of retrieval experiments when relevance judgments per se cannot be consistently assessed by human judgment would seem to represent overly vigorous pursuit of a solution before identifying the problem." Similarly, see Black, 1963 [64], p. 14: "Finally, when one is faced with an existing collection of indexed materials, how does one assess the effectiveness of any retrieval system? Suppose that one receives 20 documents as a result of a query to the system. Suppose further that all 20 documents are quite pertinent to the topic of interest. Is there any way to assess the amount of pertinent information still unretrieved from the file? Or is there any way of learning whether the retrieved information is more pertinent than the unretrieved information ? The answer is 'No!' -- the use of any retrieval system is, then, an act of faith in the quality of indexing."

7.1 Core Problems

First and foremost of the core problems implicit in the question of evaluation of any indexing scheme, whether applied by man, machine, or man-machine combinations, are those of interpersonal communication itself, which in turn relate to fundamental problems of epistemology. These are, first, the problems of language as a means of communicating perceptions, apperceptions of relationships between present observations and prior experience, and value judgments based thereon, and, secondly, even more fundamentally, the question and the veridicality of language representations of real transactions and events. Serious investigators in the field, including many who have themselves contributed to automatic indexing techniques, have made such typical acknowledgments of the difficulties as the following:

> "The imprecision connected with discussion of retrieval effectiveness and of relevance is not due to lack of understanding of the relatively straightforward retrieval processes, but is due to our lack of basic understanding about language, meaning and human communication itself." [1]

> "Fundamentally, the study of inquiry procedures is a problem in the general psychology of cognitive functioning. Relevant problems concern the way problems are recognized and formulated into questions, the way a search plan is developed to find answers to questions, and finally, the way it is decided whether or not a possible answer matches the specifications of a question." [2]

A second core problem is the heterogeneous and somewhat arbitrary development of natural languages themselves. It is much the same fundamental problem whether men or machines are to read text and determine the "meaning" (at least, in the sense of communication intent) of messages expressed in a natural language. However, the problems are aggravated if men themselves must know enough about language and its conveyances of message content to specify precisely to a machine what it is to look for and to use.

Salton enumerates some of these difficulties as follows:

> "No well-defined set of rules is known by which the individual words in the language are combined into meaningful word groups or sentences. Specifically, the correct identification of the meaning of word groups depends at least in part on the proper recognition of syntactic and semantic ambiguities, on the correct interpretation of homographs, on the recognition of semantic equivalences, on the detection of word relations, and on a general awareness of the background and environment of a given utterance." [3]

[1] Giuliano, 1963 [230], p . 6.

[2] Stone, 1962 [576], p. 1.

[3] Salton, 1963 [519], p. I-2.

Similarly, Baxendale states:

> "We are confronted with difficulties which arise from the multiple ways in which words and sentences are put together to convey meanings and shades of meaning -- i. e., to represent ideas and concepts. Research into this problem -- drawing upon psychological and logical analysis -- is scarcely begun." [1]

A third core problem is the proper choice of appropriate selection criteria if *condensed* representations of document content must be used for scanning, search, and relevance decisions. Swanson suggests that the price paid for brevity of representation so that searching operations can be efficiently managed is the loss of at least some, perhaps most, of the information in a collection or library. He notes also that:

> "It is another obvious but seldom remarked fact that the extent of such information loss for existing libraries is not only unknown but has never defined in measurable terms." [2]

This loss is lived with, today, in many practical situations involving abstracts, index term sets, selective-dissemination notices, and even mere author-title listings in announcement bulletins or search output products from either manual or machine searches. Yet the sheer increase in volume of the total number of items to be covered and of the number of items potentially responsive even to a single individual's interests has severely stretched any individual's capacity to scan or skim, much less read, the presumably pertinent material -- documents themselves, abstracts of other documents, listings of documents available -- already accumulating on his desk.

Condensation, reductive representation, becomes more and more imperative. Concurrently, while conventional tools may be lived with, after a fashion, the substitution of machine-compiled or machine-produced alternatives, even though they give the same information in the same volume, number of pages to be scanned, may because of such things as inferiorities of page and line formatting, size of type on the page, limitation of typography to upper case and a few other symbols, make the problem of how adequate the user judges the selection and condensation to be, that much worse.

A fourth problem in evaluation, therefore, is the question of whether or not the benefit to users is worth the cost. For example, despite the arguments for concept rather than word indexing, for assignment of labels rather than mere extraction of a few words used by the author himself, at least some data on the use made by scientists of various sources of information on material which might be of interest to them suggests

[1] Baxendale, 1962 [42], p. 68.

[2] Swanson, 1960 [582], pp. 5-6.

that subject indexes are not the most important source, nor even a major source. Herner found, for example, that only about 16 p ercent of his respondents reported use of indexes and abstracts as primary tools in literature searches. He reports, for the use of tools in becoming aware of current sources of information, 477 of 3832 responses indicating the use of indexing and abstracting publications as against 486 using footnotes or other cited references, [1] 291 using library acquisition lists, and 212 using separate bibliographies (Herner, 1958 [265]).

These data, and similar findings of Fishendon that 17 percent of scientists queried considered the scanning of titles in accession lists and announcement bulletins a principal means to find information of interest, [2] suggest that KWIC type indexes may be adequate for many purposes. On the other hand, the KWIC index to the U. S. Government Research Reports made available to the public on an experimental basis through the Office ot Technical Services was discontinued after a year of subsidized operation because too few of the users indicated willingness to pay a fee in order to have the service continued on a subscription basis.

The evaluational problem here involves the lack of information on indexing costs, the relatively few quantitative and objectively validated studies that have been made of user needs, the question of whether what the user says he does or wants is what he really wants or does, and the matter of defining "interest" for different users with differing purposes and requirements. The concept of "interest" is taken to mean the motivations of a particular user or group of users at a particular time, while the equally imprecise notion of "relevance" refers to the value judgments made by the user as to the relation of an item to his query or interest.

A final core problem, then, is that of the question of relevancy itself, involving recognition that "relevancy is a comparative rather than a qualitative concept ... (and) ... that a document of little relevancy in the eyes of X might well be highly relevant in the eyes of Y." [3] Mooers states, similarly, that:

> "There is no absolute 'Relevance' of a document. It depends upon the person and his background, the work and the date. What is not relevant today may be relevant tomorrow." [4]

Good discusses various possible measures of 'relevance' - logical measures, frequency measures, references to, citations of, interest measures, linguistic measures, [5]

[1] Note that Herner's data and those of Glass and Norwood, 1958 [232], reporting 6.9 percent use of cross-citations in another paper as the method of learning of important work as against 1.2 percent using an indexing service, appear to re-enforce the claims of those who advocate citation indexing.

[2] Fishenden, 1958 [197], p. 163.

[3] Bar-Hillel, 1959 [33], p. 4-8.4.

[4] Mooers, 1963 [423], p. 2.

[5] Good, 1958 [234], pp. 7-9.

but except for the obvious statistical criteria, the problems of how to measure relevancy remain largely unresolved.

At least some data on the variability of relevance judgments is available in reports of the performance of an SDI (Selective Dissemination of Information) system. In such systems, the indexing terms or tags assigned to a new item are compared with a file of "user-profiles" that is, with a pre-prepared listing of terms or topics in which a particular user is interested. Where the term-profile of a new item matches that of a user, a notification of the acquisition of that item is sent to him. Barnes and Resnick report tests of such a system in which pseudo-notifications selected randomly were included with those produced from the matching procedure. Account was kept of which notices were regarded by the users as meeting their interests and which were not. They found that 58.1 percent of the non-random notifications were regarded as relevant, but that so also were 26.8 percent of the random ones. [1/]

Katter comments on findings that the intersubjective agreement of typical users with respect to value judgments of condensed representations of text is low. He suggests:

> "One source of this low intersubjective agreement among users may be that it is often not clear what is intended by the words relevant and representative. Considerations such as the validity of the material, its usefulness, stylistic qualities, understandability, conceptual preferability, etc., can all enter their judgments in unknown amounts." [2/]

Corroborating evidence is available from other sources. Swanson, in his tests of a natural language text searching technique, had first used subject matter specialists to rate the relevance of each of the text documents to each of 50 questions. Two individuals rated each item, and if they disagreed significantly, a third person was asked to reconcile the difference. In spite of this, 8 percent of the cases of failure to retrieve "relevant" documents were ascribed to incorrect initial judgments of relevance, and 15 percent of the presumably "irrelevant" documents were finally judged to be relevant after all (Swanson, 1961 [586]). In Swanson's words: "The question of formulating criteria for judging the relevance of any document to the motive, purpose, or intent which underlies a request for information is profound and lies at the heart of the matter." [3/]

[1/] Barnes and Resnick, 1963 [36], p. 2.

[2/] Katter, 1963 [308], p. 24.

[3/] Swanson, 1960 [587], p. 1099.

7.2 Bases and Criteria for Evaluation of Automatic Indexing Procedures

What should the bases be for the evaluation of existing or proposed indexing systems that rely, to a greater or lesser extent, on machine generation of the indexing or classificatory labels? Since the evaluation of quality of indexing per se raises such fundamental and elusive questions, can these questions be begged for the case of automatic indexing as they are in fact for almost all manual systems? If so, the obvious bases are those of time, cost, availability of alternative possibilities, and customer acceptance. Here again we are faced with a dearth of objective data, even for the intercomparison of any two manual systems.

In the two years preceding the ICSI Conference, the Program Committee openly solicited papers that would provide comparative data for operating information systems and that would develop and discuss criteria for the comparison of systems. [1] Nevertheless, of the papers received only two were responsive to this invitation: the special case of comparing the conventional file against the inverted file approach to the searching of chemical structure data (Miller et al, 1959 [419]), and an early report by Cleverdon on the ASLIB Cranfield project for the intercomparison of indexing systems, under a grant from the National Science Foundation (1959 [126]).

There had been an earlier comparative experiment, generally conceded to be the first of its kind, [2] in which 98 search requests were run by ASTIA personnel using a conventional catalog and by personnel of Documentation Inc., using a coordinated uniterm index. Warheit says:

> "Unfortunately, the conditions of the test were very poorly designed so that, in the final analysis, each group was the sole judge both of the scope of the original request and of the adequacy of the bibliographies produced. The resulting claims are of course contradictory." [3]

[1] See "Proposed Scope of Area 4," Proceedings, ICSI, 1959 [481], pp. 665-669.

[2] Compare, for example, Gull, 1956 [246], p. 329: "When one considers that a fairly thorough search of the literature indicates that this comparison of two reference systems is the first undertaken so far, it is not surprising that the results reveal clerical errors and an incomplete design of the test."

[3] Warheit, 1956 [631], p. 274.

However, some of the findings are pertinent to our present questions of evaluation. Thus, of 492 items selected by Documentation, Inc., that ASTIA considered pertinent but had not selected, 98 were missed by them although the proper subject heading was searched and the catalog card had adequate selection clues, 89 were missed because not all applicable subject headings were searched, 21 were missed because the original subject heading assignments had been inadequate, 7 were missed because neither title nor abstract provided indication that the report itself was pertinent to the request, and 102 were missed "because the subject heading did not occur to the searcher or because there were so many cards under the subject heading that the searcher was discouraged". [1/] Similarly, Gull reports, of 318 items selected by ASTIA that Documentation, Inc. personnel considered relevant but had not themselves selected, 97 were missed because the searcher did not consult the proper terms.

7.2.1 The Cranfield Project

The inauguration of the Cranfield project is itself indicative of a prior lack of objective standards as applied to the measurement of effectiveness of information indexing, selection and retrieval systems. [2/] Beginning in 1957, and still continuing with respect to individual indexing devices such as synonym controls and role indicators, this work has attempted to compare different indexing systems (e.g., UDC, Uniterm, etc.) under different indexing conditions (e.g., type of training of indexer, length of time allowed to index) against proposed measures of "retrieval effectiveness". These measures are, respectively, the recall ratio, or the percentage of relevant documents retrieved as against the total number of relevant documents known to be in the collection, and the relevance ratio, or the percentage of relevant documents among those actually retrieved.

In the first Cranfield tests, on 18,000 documents, it is reported that the recall ratio ranged between 75 and 85 percent for all four indexing systems. [3/] These results are

[1/] Gull, 1956 [246], p. 329.

[2/] Compare, for example, Randall, 1962 [492], pp. 380-381: "Prior to 1957, the proponents of the various indexing and classification schemes, the universal decimal system, the alphabetic subject heading, the Uniterm system and faceted classification touted their own system on the bases of subjective evaluation and theoretical investigations. There were many claims and much supposition about the relative merits and benefits ... but there was no body of data from which an objective evaluation could be made... Many observers believe that the Cranfield study constitutes the most important work done in the field of cataloging in recent times."

[3/] Cleverdon, et al, 1964 [130], p. 87.

rather better than reported by others [1] and have been subjected to specific criticisms although these first tests were limited to the recall of the source documents on which the test questions were based. For non-source documents there would of course also be questions relating to the core problem of how relevance is to be judged. Thus Markus says:

> "Despite investigations by Cleverdon in England, and by many others, there is today no generally accepted method of comparing the effectiveness of different types of indexes. The needs of index users vary so greatly that even the most carefully planned tests of retrieval efficiency can be challenged." [2]

Notwithstanding such criticisms, however, and in spite of the fact that the Cranfield tests have so far been directed principally to indexing systems applied manually, certain findings and conclusions reached by Cleverdon and his associates are pertinent to the questions of evaluating automatic indexing procedures. Examples are:

> "The fact is that no indexing sleight of hand, no indexing skill, can produce a system in which a figure for recall can be improved substantially without weakening the over-all relevance, i.e., the number of documents that are really relevant compared with the total number retrieved.
>
> "The majority of the failures (60 percent) were due to inadequacies and inaccuracies (carelessness rather than lack of knowledge) in the indexing process. However, supplementary tests, in which the staff of outside organizations carried out the indexing revealed that the Cranfield indexers were achieving a standard above average. This seems to indicate a certain inevitability of human weakness and error in the indexing process and lends some support to the many current research projects that are investigating the feasibility of automatic indexing." [3]

7.2.2 O'Connor's Investigations

As O'Connor has cogently observed on a number of occasions, the question of whether or not automatic indexing is possible is not the real question. Rather, the problem is whether or not indexing by machine is capable of producing results that are "good enough" for retrieval purposes, raising in its turn the still more basic question of how "good retrieval" can be evaluated. His own approach in detailed investigations has

[1] See, for example, Johnson 1962 [300], p. 90: "The amount of meaningful information that can be retrieved is too small. There are few available studies on this subject. But these seem to indicate that, under some indexing schemes, meaningful retrieval can run as low as 10 and 15 percent and that the most that can be optimized for any of them, even under highly motivated conditions, is around 70 percent."

[2] Markus, 1963 [394], p. 16. See also Kochen, 1963 [327], p. 12: "The outstanding large-scale and realistic experimental work is that of Cleverdon. Unfortunately, his results are not very decisive."

[3] Cleverdon et al, 1964 [130], pp. 86-87.

been to study an existing system (e.g., using Merck, Sharp and Dohme data) with respect to indexing terms such as "penicillin," "toxicity," and "mode of action." He then attempts to define various possible machine assignment rules, and then to determine the probable over-and-under assignments that would result from the application of these rules.

Typical results pertinent to both questions of word-indexing evaluation and of inter-indexer consistency showed that for 23 documents indexed under the term "toxicity," 11 did not contain the stem "toxi..." at all; that 17 items indexed under "penicillin" contained the word at least once; that none of 34 randomly selected documents not indexed under "penicillin" contained the word, but that 7 of 28 items not so indexed but selected as probable candidates from title and other clues did contain the word. (O'Connor, 1961 [447])

Typical suggestions, comments, and conclusions made by O'Connor include the following:

> "It might be required that the mechanized indexing permit as good (or no worse) retrieval as existing human indexing, because it is desired to free the subject-skilled indexing personnel for other work. Or poorer retrieval (than possible with human indexing such as is presently done of comparable material) might be accepted from computer indexing, because poorer retrieval is better than none and there is a shortage of subject-skilled people to do the additional indexing." [1]

> "Such considerations as the following are relevant. Over-assigning can increase input costs and storage (to an extent dependent on the storage system), but mechanizing indexing might be worth the cost. Over-assigning might also increase the number of irrelevant documents retrieved, but the increase might be insignificant." [2]

> "...Suppose terms A, B, and C each correctly characterize five percent of a ten thousand document collection, each term is overassigned to another five percent, and over-assignment of each term occurs independently of the correct assigning and over-assigning of the others. Then about nine documents will be extra for the search question A & B & C." [3]

> "The question of permitting some under-assigning, that is, the computer failing to assign [a term] T to some document which should have it, is more delicate. Human indexers sometimes underassign. If we knew the rate of ounderassigning by human indexers for a term T, we might consider allowing the computer a similar rate. However, some cases of underassigning might be more important than others and if the computer made more important mistakes than the human indexers, retrieval might not be 'good enough'." [4]

[1] O'Connor, 1960 [444], p. 3.

[2] O'Connor, 1961 [448], p. 199.

[3] O'Connor, 1960 [444], p. 6.

[4] Ibid, pp. 6-7.

Other typical points made by O'Connor include the possibilities that the use of automatic indexing techniques might free trained technical people for other work, that it might permit more indexing than is now possible with available resources, that it might cost less, and that it might produce a better or more consistent indexing product.[1] With respect to the latter point, however, he points out that greater consistency might not in itself be a virtue, since the product although generated more consistently might be relatively worthless by comparison with the inconsistent human product. [2] Especially pertinent to the question of judgment factors in evaluation was a comparison of the most frequent words selected by the Luhn "auto-encoding" technique as applied to an ICSI paper against a quasi-random word list for the same paper produced by selecting the last non-common word on every page, and the first such word on every second page. He remarks:

> "The important point of this quasi-random list for my present purposes is to emphasize that first impressions might not be at all a good way of judging the adequacy of an index set." [3]

7.2.3 Questions of Comparative Costs

The paucity of objective data on the effectiveness of indexing systems generally extends to even such obvious questions as costs of indexing and time required to index. These very questions might, in fact, be decisive with respect to choice between manual and machine systems. It has been estimated by some that the costs of manual subject indexing amount to close to 75 percent of the costs of operating an information selection and retrieval system, [4] yet very little actual data on costs has been reported in the literature. [5] Exceptions are, for the most part, limited to rather special cases, such as the following examples:

1. A total cost of less than $30,000 is reported for a 10,000 document collection at Aeronutronic. Four man-years of effort were required. On average, 12.6 access points were provided per document, of which 9.2 were subject-indicating descriptors chosen, with some modifications, from the second Edition of the ASTIA Thesaurus. "This favorable figure was possible because an adequate ready-made thesaurus of indexing terms was available and because the 'peek-a-boo' type equipment used was much

[1] O'Connor, 1962 [447], p. 267.

[2] O'Connor, 1963 [443], p. 16.

[3] O'Connor, 1962 [447], p. 270.

[4] O'Connor, 1963 [442], p. 1.

[5] See, for example, A. D. Little, Inc. (1963 [23], p. 5): "Performance and cost data on existing large documentation systems are surprisingly sparse, and cost data have rarely included adequate overhead and depreciation accounting."

less expensive than most other devices offering comparable speed of operation and search logic possibilities." [1]

2. "The experience of libraries that have gone through indexing using links and role indicators and careful editing shows that indexing takes about one-half hour per document (or $4.00) and costs an additional $1.00 for routine processing." [2]

3. In an investigation of the comparative merits of manual indexing of 2,000 documents using the UDC classification system as against a KWIC index, Black gives the figure of approximately $1400 for the UDC case compared to about $600 for an in-house computer operation to produce KWIC listings, and somewhat more for a KWIC index compiled by a service bureau. [3]

Time required to index, which directly involves cost, is reported by Cleverdon to vary widely:

> "Few reliable figures have been given for current practices, although a particularly high figure is the 1 1/2 hours average quoted for indexing reports for the catalogue of aerodynamic data prepared by the Nationaal lluchtvaart laboratorium in Holland. It appears from personal discussions that an average of 20 minutes for a general collection of technical reports is the top limit, and this has been taken as the maximum indexing time to be used in the project." [4]

Insofar as such meagre data is indicative, there does not appear to be any particular cost-advantage for machine-compiled and machine-generated indexing other than the title-only KWIC indexes. Thus, Olmer and Rich report, in part:

> "The program ... lends itself to a variety of applications. One of these ... is estimated to cost roughly $4.00 per document for cataloguing, putting on tape, printing and making any necessary corrections." [5]

This is for a case where the indexing (cataloging) is done manually.

For a specific proposed automatic indexing system, employing a modified version of the Luhn word-frequency counting selection principle, Gallagher and Toomey report that:

[1] Linder, 1963 [361], p. 147.

[2] Lockheed Aircraft Corp., 1959 [369], p. 93.

[3] Black, 1962 [65], p. 318.

[4] Cleverdon, 1959 [126], p. 690.

[5] Olmer and Rich, 1963 [454], p. 182.

> "For the documents in our system, we estimate that processing time will be about 20 seconds per thousand words ... The cost is approximately $3.50 per minute when averaged between prime and extra shift." [1]

This means that the cost of processing a 3,000-word document would be $3.50 , exclusive of the costs of keypunching the input text which, conservatively estimated, costs not less than 1-2 cents per word. [2] Swanson similarly assumes either that machine-usable text is already available or that editing and keystroking efforts are separate costs in arriving at an estimate of $1.00 per item for automatic indexing. [3]

These quantitative estimates bear out the more subjective conclusions of such investigators as Bar-Hillel, O'Connor, and others. Examples are:

> "It is very likely that manual uniterm indexing by cheap clerical labor will still, on the average, be qualitatively superior to any kind of automatic indexing, and it is very unlikely that the cost of automatic indexing will ever be less than this kind of manual uniterm indexing, unless the automatic indexing is to be of such low quality as to totally defeat its purpose." [4]

> "Most of these techniques require that the full texts of documents be in machine readable form. At present this usually requires keypunching which is much more expensive than a specialist's indexing efforts." [5]

[1] Gallagher and Toomey, 1963 [205], p. 52.

[2] 'Compare, for example, Ray, 1961 [496], p. 55; Swanson, 1962 [584], p. 470: The cost is roughly one or two cents per word which by standards of what is normally spent for even the most thorough indexing and cataloging, is exorbitant." Mersel and Smith report 1964 [415], p. 10A) typical TRW costs of keypunching as two cents per word for Russian technical text, and one cent per word for English. They also cite cost figures as low as half a cent per word at the CIA-Georgetown Keypunching Center in Frankfurt and at IBM, but this is exclusive of overhead and computer processing (e.g., editing program) costs, so that the one cent figure appears minimal as of today. However, Kochen reports (1963 [327], p. 7): "While keypunching of text cost roughly one cent/word, new means for recording spoken (and written) text using a steno-keyboard tied to a photodisc storing a Stenocode-English dictionary could possibly reduce the cost to 1/3-cent per word."

[3] Swanson, 1962 [584], p. 471.

[4] Bar-Hillel, 1962 [35], p. 418.

[5] O'Connor, 1963 [443], p. 1.

7.2.4 Summary: Potential Advantages as Bases for Evaluation

In view of the difficulties engendered by the underlying core problems, the criticisms that can be brought against tests of "retrieval effectiveness", the general lack of comparative data and standards of measurement, the question of evaluation of automatic indexing procedures largely reduces to the weighing of potential advantages and disadvantages. In the case of such procedures as KWIC and citation indexing, some of these possibilities, both pro and con, have been discussed previously. In general, suggested bases for evaluation reflecting operational considerations may be summarized as follows:

1. Speed and timeliness
2. Relative economy
3. Consistency and reliability[1/]
4. Elimination of the need for further human intellectual effort after initial planning and programming has been done.
5. Providing a product that could not otherwise be obtained.
6. Ease of updating and revision of indexes so produced.[2/]

From the point of view of possible operational advantages, these may be combined into the single criterion:

> The achievement of a more effective and more economical balance between the meeting of the objectives of the indexing system and the utilization of available resources.

[1/] Compare McCormick, 1962 [409], p. 182: "A computer is objective in its operations and it can be repetitive. If given a certain amount of information about a document, it is always able to index the document in a consistent manner. This consistency is desired so as to avoid the situations where a person might index a document differently on various occasions, or where it would be indexed differently by another person when there appears to be no good reason for a difference." Note, however, O'Connor's point previously mentioned, (1963 [443], p. 16): "It has been argued that mechanized indexing has the advantage of consistency... However this argument by itself says very little in favor of mechanized indexing. For two humanly produced index sets for a document which differ somewhat may both be quite useful, though imperfect, while the index set which the same program will always reproduce for the same document may be worthless."

[2/] See, for example, Youden, 1963 [658], p. 332: "The facility with which indexes may be updated and the ease of selecting items for special bibliographies will result in the majority of indexes being computer produced before many years."

However, the question of the objectives of the system brings us back full circle to the questions of purpose in terms of particular requirements, of quality, and of how to measure either purpose or quality. Thus we may determine that an automatic indexing procedure produces a product at least as rapidly, at least as inexpensively, at least as consistently as human indexing operations would, and with substantially less investment of manpower resources. However, will this product be as useful or as "good" as the human product?

In view of the many caveats about the present quality of indexing systems[1] and the lack of standards for measuring quality, [2] it is important to recognize that we should compare the products of automatic indexing methods "not with hand-crafted excellence, but with the average, the routine output of the over-burdened subject analyst working with the deficiencies of any other indexing system". [3] Such deficiences include the critical question of how well and how consistently the system, whatever it is, is applied in practice by the human analysts.

7.3 Findings with Respect to Inter-Indexer and Intra-Indexer Consistency

Very few objective studies, despite the obvious relationship to the general questions of quality, pertinency, and reliability of indexing, have as yet been made of inter-indexer and intra-indexer consistency. Perhaps the first investigation both to obtain experimental data and to analyze the observed types of failures to achieve correct assignments was that of Lilley. [4] He took the answers made to 6 questions by 340 students entering a graduate library school, wherein they were asked to write down the subject headings which they would expect to be applied to other books on the same subject as 6 "sample books" in a system such as the Library of Congress card catalog. Lilley reports:

[1] See, for example, in addition to comments by O'Connor and others previously quoted, Helyar, 1961 [262], p. 110: "The general current of feeling of the meeting as reflected both in the papers and in the discussion is that the standard of indexing is not nearly adequate;" Artandi, 1963 [22], p. 1.: "... 'Good indexing' as such has not been defined satisfactorily and is the function of many variables, some known, others not yet identified"; Tritschler, 1963 [610], p. 5: "... 'Good'indexing is extremely difficult to describe and 'perfect' indexing is impossible to define or measure."

[2] See Cleverdon, 1960 [124], p. 429: "The most important requirement in information retrieval is a recognized standard of measurement and after that we need a satisfactory method of measuring. Only when these have been found will it be possible to know for certain whether any new system of indexing or retrieving information is an improvement on previous methods. At present all those trying to solve the problems of information retrieval are working very much in the dark, uncertain as to the real problems and quite unable to apply any measurements to their proposed solutions."

[3] Kennedy, 1962 [311], p. 126.

[4] Lilley, 1954 [360]: See also Vickery, 1960 [626], p. 4.

> "A total of 2245 headings were suggested, averaging 1. 1004 headings per book per student. These headings represented 373 different varieties, of which 368 were different from the headings traced on the Library of Congress cards for the sample books... As an average 62. 17 different headings were suggested for each book...
>
> "When the 368 different varieties of incorrect headings were analyzed in accordance with certain criteria that had been set up, it was found that incorrect specificity was a factor in 93. 48%, incorrect terminology in 79. 08% and incorrect form of entry in 72. 28% of the headings... Over half of the incorrect headings (54. 62%) had some combination of two errors, and almost half (49. 73%) could have been converted into 'correct' headings only by changing the level of specificity, and by revising the terminology, and by altering the form...
>
> "It was also found, contrary to the general assumption that failure in specificity almost always means that the reader is approaching his subject from too broad a point of view, that of those headings in which an incorrect level of specificity was a factor... 64. 82% were too broad and 35. 18% were too narrow." [1]

Lilley then asks the rather plaintive question as to what would happen, given that his quite homogeneous group of subjects, all of them college graduates and all seriously interested in librarianship, could come up with more than 62 different headings, on average, for every heading actually used in the catalog, if his test group had included a larger number of subjects with more heterogeneous interests?

In 1961, Macmillan and Welt investigated the duplicate indexing of 171 papers in a limited area of the medical sciences (1961 [389]). In only 18 percent of the cases was the indexing identical or nearly so. About a third of the papers had been indexed so differently that there was no common correlation. For the rest, terms were used in one case that were missed in the other.

Some brief data on inter-indexer consistency is also provided by Kyle (1962 [342]) for two indexers applying her classification system to 246 arbitraily selected French and English items in the field of political science. Of these, 160 were indexed the same way by both indexers, for a consistency figure of 70 percent. Tritschler noted that no items were indexed the same way a second time as they were the first, in small-scale experiments involving 20 documents independently indexed by 7 different people. [2]

Painter (1963 [460]), in her study of problems of duplication and consistency of subject indexing of the reports handled by the Office of Technical Services, proceeded by selecting items from the announcement bulletins of agencies contributing to OTS, having these items re-indexed in the various agencies, and comparing the results with the original indexing assignments. At ASTIA, 94 items were re-indexed, with 1, 239 terms having been assigned to them originally and 1, 119 assigned on the re-run. Overall, 62 percent of those terms originally assigned were also assigned the second time, and 69 percent of the second-time terms had also been assigned originally. However, 111 of the starred descriptors (which are of the most significance in the ASTIA system) were used the first time and not the second, while 98 were used the second time but not the first.

[1] Lilley, 1954 [360], pp. 42 and 43.

[2] Tritschler, 1963 [610], p. 5.

At AEC, 96 items were re-indexed to the subject heading scheme used in Nuclear Science Abstracts. There had been 249 headings assigned to these items originally and 406 were assigned on the second run, for an overall consistency rate of 54 percent, but with 53 percent of the headings used the second time not having been used the first. The sample checked at OTS consisted of 32 items to which 346 descriptors had been assigned the first time and 418 the second. The consistency was 65 percent with respect to the first run and 54 percent with respect to the second. Finally, at the National Agriculture Library 99 items were checked, with results showing a high consistency rating and a similarity of indexing between the two runs of 86 percent. Painter concludes:

> "The consistency rates are not encouraging. Apparently there is little difference between preparation for a manual system and that for a machine system. The percentages indicate that there is no significant difference between consistency where two or three headings are assigned and where twelve or sixteen are assigned. Therefore, we are left with the fact that regardless of these variables, consistency rates range between 60 and 72 per cent." 1/

Jacoby and Slamecka report even less encouraging data (1962 [293]). "In general, the inter-indexer reliability was found to be low (in the vicinity of 20 per cent), the intra-indexer reliability somewhat higher (about 50 per cent)." For a series of tests of indexing of a group of chemical patents by three experienced and three inexperienced indexers, they found that the beginners had average matchings among the terms assigned by them to the same documents of only 12.6 percent and that even for the experienced indexers the average percent of matching terms was only 16.3 percent. 2/ In other studies, these investigators have explored the effects of various indexing aids upon the reliability and consistency of indexing, concluding that the use of prescriptive aids such as authority lists improves reliability and inter-indexer consistency from 8 or 9 percent to 33 percent, while those aids such as thesauri and association lists "which enlarge the indexer's semantic freedom of term choice" are detrimental (Slamecka and Jacoby, 1963 [560]).

Rodgers in a study of intra-indexer consistency reports data for the re-indexing, by the same person at a later date, of 60 documents dealing with the United Arab Republic taken from The New York Times. She reports that the average consistency over all 60 documents was 59 percent. 3/ In a further study of inter-indexer consistency, 20 papers from Area 5, ICSI, were key-word indexed by 16 people all of whom were familiar with the subject matter, (although only 8 completed all 20 papers). Results are given in terms of the proportions of the total number of unique words chosen by 100 percent of the subjects (.008) half of them (.14) and only one of them (.52). 4/ Study of the results in terms of the proportion of words selected in common by any pair of these indexers to the total number of different words selected by them both gave a "grand mean agreement for all two-person combinations for the 8 subjects... [of].. 24 percent against all 20 articles." 5/ The mean percentage of overlap between Luhn's word-frequency selection technique (as applied to the same papers) and any one or more indexers who agreed was .15.

1/ Painter, 1963 [460], p. 94.

2/ Jacoby and Slamecka, 1962 [293], p. 16.

3/ Rodgers, 1961 [504], p. 12.

4/ Rodgers, 1961 [503], p. 50.

5/ Greer, 1963 [239], p. 10.

Still further studies of indexer consistency investigated at the Information Systems Operation division of General Electric have just recently been reported (Korotkin and Oliver, 1964 [331, 332]). In particular, the investigators report on the effects of subject matter familiarity and on the use as a job aid of a reference list of suggested descriptors upon inter-indexer consistency. The material for test consisted of 30 abstracts drawn from Psychological Abstracts, to be indexed by 5 psychologists and 5 non-psychologists in two sessions, with and without use of the "job aid". Results in terms of mean percent consistency were reported as follows:

	Session I	Session II
"Group A (Familiar)	39.0%	53.0%
Group B (Non-familiar)	36.4%	54.0%" 1/

Corroborating evidence of a generally low rate of inter-indexer consistency is provided by noting instances of duplicated indexing that may occur in regularly issued announcement bulletins. During current awareness scanning of the DDC (ASTIA) "TAB" in recent months, members of the staff of the Research Information Center and Advisory Service on Information Processing have caught more than 20 cases of duplicate and even triplicate indexing of the same item. (Two examples can be discovered in Figure 8 a and b). For the 52 independent assignments involved, for these items the average inter-indexer consistency is only 46.1 percent.

On the general subject of indexing consistency, Black comments as follows:

> "There have been enough experiments to indicate that there is no consistency, or very little, between one indexing performance by a given individual and another indexing performance, at a later date, by the same individual. The same inconsistency has been discovered among different individuals all indexing the same documents. Thus there is neither inter-indexer consistency nor intra-indexer consistency in any system that depends on human performance." 2/

There can be little doubt that the quality and consistency of most human indexing, practically available today, is not good. Much of it, because of time and other pressures, is either directly a word-extraction process, or it is inconsistent in assignment of many relevant descriptors and subject category labels. On the other hand, today's indexing, whether accomplished by man or machine, is probably no better and no worse than any other classificatory or indexing procedures. The only excuse, therefore, for choice between man and machine is the cost/benefit ratio which is related on the one hand to specific operational considerations and on the other to the question of whether or not various indexers, and various users, would agree with the machine as much as they agree with each other.

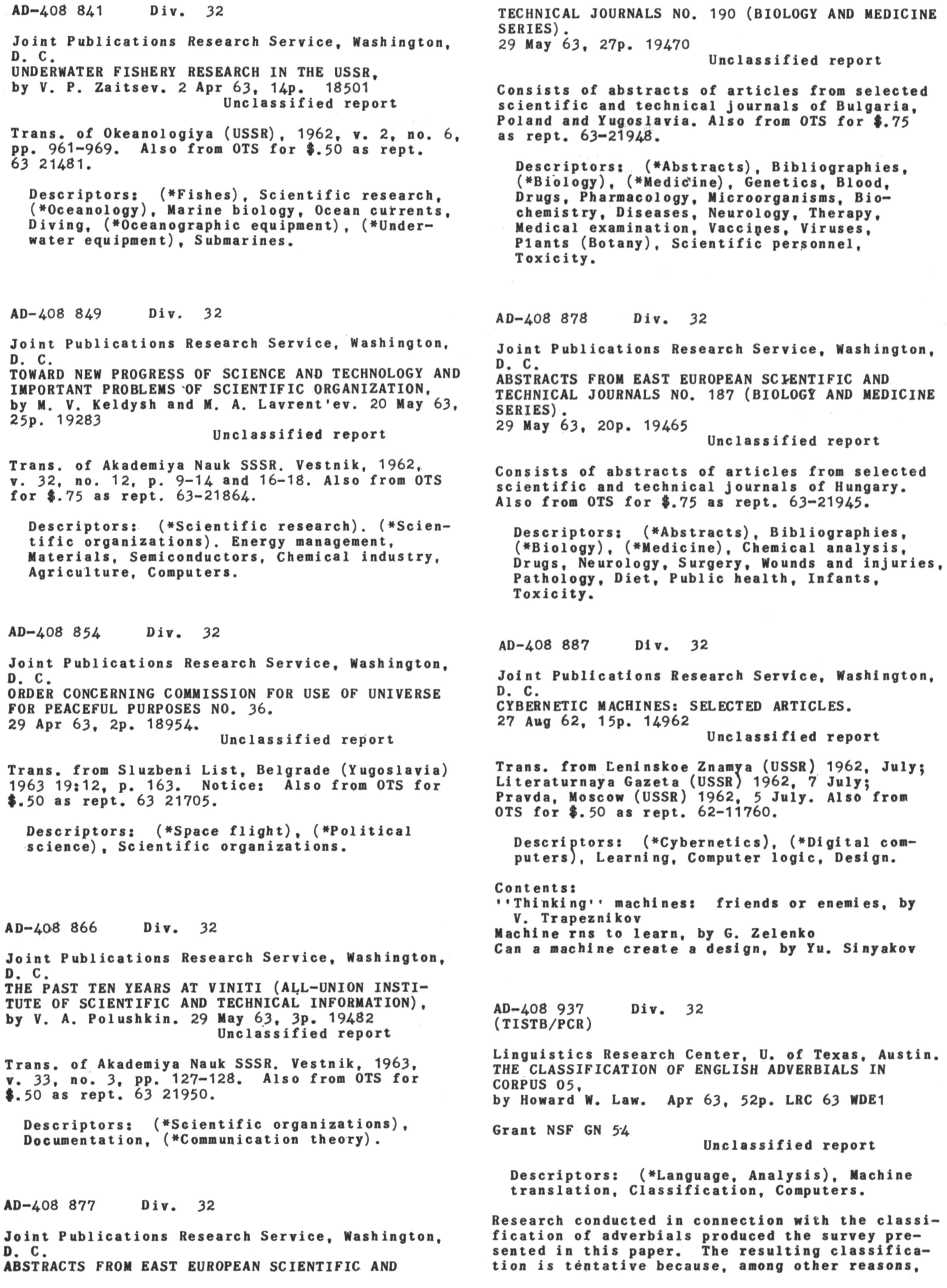
Before turning to some of the operational considerations affecting the cost-benefit ratio, however, certain special factors should be briefly mentioned.

7.4 Special Factors and Other Suggested Bases for Evaluation

The difficulties and problems of evaluation so far considered are generally applicable to any indexing system, whether manual or automatic. Certain special factors arise, however when we consider some of the proposed automatic assignment and automatic classification techniques. In addition, the prospects for computer processing hold at least the

1/ Korotkin and Oliver, 1964 [331], p. 7.

2/ Black, 1963 [64], pp. 16-17.



AD-408 841 Div. 32

Joint Publications Research Service, Washington, D. C.
UNDERWATER FISHERY RESEARCH IN THE USSR,
by V. P. Zaitsev. 2 Apr 63, 14p. 18501
Unclassified report

Trans. of Okeanologiya (USSR), 1962, v. 2, no. 6, pp. 961-969. Also from OTS for $.50 as rept. 63 21481.

Descriptors: (*Fishes), Scientific research, (*Oceanology), Marine biology, Ocean currents, Diving, (*Oceanographic equipment), (*Underwater equipment), Submarines.

AD-408 849 Div. 32

Joint Publications Research Service, Washington, D. C.
TOWARD NEW PROGRESS OF SCIENCE AND TECHNOLOGY AND IMPORTANT PROBLEMS OF SCIENTIFIC ORGANIZATION,
by M. V. Keldysh and M. A. Lavrent'ev. 20 May 63, 25p. 19283
Unclassified report

Trans. of Akademiya Nauk SSSR. Vestnik, 1962, v. 32, no. 12, p. 9-14 and 16-18. Also from OTS for $.75 as rept. 63-21864.

Descriptors: (*Scientific research). (*Scientific organizations). Energy management, Materials, Semiconductors, Chemical industry, Agriculture, Computers.

AD-408 854 Div. 32

Joint Publications Research Service, Washington, D. C.
ORDER CONCERNING COMMISSION FOR USE OF UNIVERSE FOR PEACEFUL PURPOSES NO. 36.
29 Apr 63, 2p. 18954.
Unclassified report

Trans. from Sluzbeni List, Belgrade (Yugoslavia) 1963 19:12, p. 163. Notice: Also from OTS for $.50 as rept. 63 21705.

Descriptors: (*Space flight), (*Political science), Scientific organizations.

AD-408 866 Div. 32

Joint Publications Research Service, Washington, D. C.
THE PAST TEN YEARS AT VINITI (ALL-UNION INSTITUTE OF SCIENTIFIC AND TECHNICAL INFORMATION),
by V. A. Polushkin. 29 May 63, 3p. 19482
Unclassified report

Trans. of Akademiya Nauk SSSR. Vestnik, 1963, v. 33, no. 3, pp. 127-128. Also from OTS for $.50 as rept. 63 21950.

Descriptors: (*Scientific organizations), Documentation, (*Communication theory).

AD-408 877 Div. 32

Joint Publications Research Service, Washington, D. C.
ABSTRACTS FROM EAST EUROPEAN SCIENTIFIC AND TECHNICAL JOURNALS NO. 190 (BIOLOGY AND MEDICINE SERIES).
29 May 63, 27p. 19470
Unclassified report

Consists of abstracts of articles from selected scientific and technical journals of Bulgaria, Poland and Yugoslavia. Also from OTS for $.75 as rept. 63-21948.

Descriptors: (*Abstracts), Bibliographies, (*Biology), (*Medicine), Genetics, Blood, Drugs, Pharmacology, Microorganisms, Biochemistry, Diseases, Neurology, Therapy, Medical examination, Vaccines, Viruses, Plants (Botany), Scientific personnel, Toxicity.

AD-408 878 Div. 32

Joint Publications Research Service, Washington, D. C.
ABSTRACTS FROM EAST EUROPEAN SCIENTIFIC AND TECHNICAL JOURNALS NO. 187 (BIOLOGY AND MEDICINE SERIES).
29 May 63, 20p. 19465
Unclassified report

Consists of abstracts of articles from selected scientific and technical journals of Hungary. Also from OTS for $.75 as rept. 63-21945.

Descriptors: (*Abstracts), Bibliographies, (*Biology), (*Medicine), Chemical analysis, Drugs, Neurology, Surgery, Wounds and injuries, Pathology, Diet, Public health, Infants, Toxicity.

AD-408 887 Div. 32

Joint Publications Research Service, Washington, D. C.
CYBERNETIC MACHINES: SELECTED ARTICLES.
27 Aug 62, 15p. 14962
Unclassified report

Trans. from Leninskoe Znamya (USSR) 1962, July; Literaturnaya Gazeta (USSR) 1962, 7 July; Pravda, Moscow (USSR) 1962, 5 July. Also from OTS for $.50 as rept. 62-11760.

Descriptors: (*Cybernetics), (*Digital computers), Learning, Computer logic, Design.

Contents:
''Thinking'' machines: friends or enemies, by V. Trapeznikov
Machine rns to learn, by G. Zelenko
Can a machine create a design, by Yu. Sinyakov

AD-408 937 Div. 32
(TISTB/PCR)

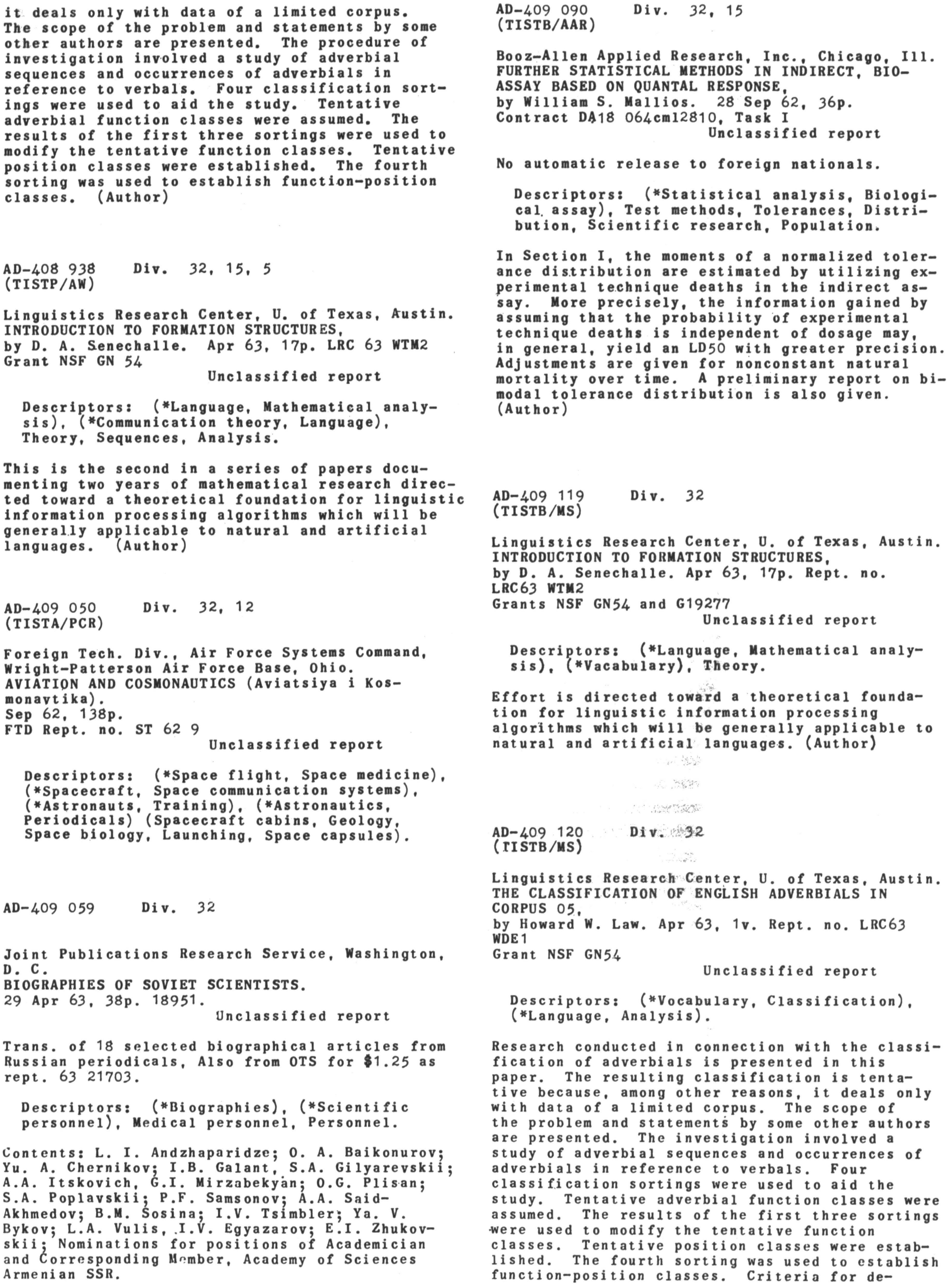
Linguistics Research Center, U. of Texas, Austin.
THE CLASSIFICATION OF ENGLISH ADVERBIALS IN CORPUS 05,
by Howard W. Law. Apr 63, 52p. LRC 63 WDE1

Grant NSF GN 54
Unclassified report

Descriptors: (*Language, Analysis), Machine translation, Classification, Computers.

Research conducted in connection with the classification of adverbials produced the survey presented in this paper. The resulting classification is tentative because, among other reasons,

Figure 8a. Examples of Duplicate Indexing



it deals only with data of a limited corpus. The scope of the problem and statements by some other authors are presented. The procedure of investigation involved a study of adverbial sequences and occurrences of adverbials in reference to verbals. Four classification sortings were used to aid the study. Tentative adverbial function classes were assumed. The results of the first three sortings were used to modify the tentative function classes. Tentative position classes were established. The fourth sorting was used to establish function-position classes. (Author)

AD-408 938 Div. 32, 15, 5
(TISTP/AW)

Linguistics Research Center, U. of Texas, Austin.
INTRODUCTION TO FORMATION STRUCTURES,
by D. A. Senechalle. Apr 63, 17p. LRC 63 WTM2
Grant NSF GN 54
Unclassified report

Descriptors: (*Language, Mathematical analysis), (*Communication theory, Language), Theory, Sequences, Analysis.

This is the second in a series of papers documenting two years of mathematical research directed toward a theoretical foundation for linguistic information processing algorithms which will be generally applicable to natural and artificial languages. (Author)

AD-409 050 Div. 32, 12
(TISTA/PCR)

Foreign Tech. Div., Air Force Systems Command,
Wright-Patterson Air Force Base, Ohio.
AVIATION AND COSMONAUTICS (Aviatsiya i Kosmonavtika).
Sep 62, 138p.
FTD Rept. no. ST 62 9
Unclassified report

Descriptors: (*Space flight, Space medicine), (*Spacecraft, Space communication systems), (*Astronauts, Training), (*Astronautics, Periodicals) (Spacecraft cabins, Geology, Space biology, Launching, Space capsules).

AD-409 059 Div. 32

Joint Publications Research Service, Washington, D. C.
BIOGRAPHIES OF SOVIET SCIENTISTS.
29 Apr 63, 38p. 18951.
Unclassified report

Trans. of 18 selected biographical articles from Russian periodicals, Also from OTS for $1.25 as rept. 63 21703.

Descriptors: (*Biographies), (*Scientific personnel), Medical personnel, Personnel.

Contents: L. I. Andzhaparidze; O. A. Baikonurov; Yu. A. Chernikov; I.B. Galant, S.A. Gilyarevskii; A.A. Itskovich, G.I. Mirzabekyan; O.G. Plisan; S.A. Poplavskii; P.F. Samsonov; A.A. Said-Akhmedov; B.M. Sosina; I.V. Tsimbler; Ya. V. Bykov; L.A. Vulis, I.V. Egyazarov; E.I. Zhukovskii; Nominations for positions of Academician and Corresponding Member, Academy of Sciences Armenian SSR.

AD-409 090 Div. 32, 15
(TISTB/AAR)

Booz-Allen Applied Research, Inc., Chicago, Ill.
FURTHER STATISTICAL METHODS IN INDIRECT, BIO-ASSAY BASED ON QUANTAL RESPONSE,
by William S. Mallios. 28 Sep 62, 36p.
Contract DA18 064cm12810, Task I
Unclassified report

No automatic release to foreign nationals.

Descriptors: (*Statistical analysis, Biological assay), Test methods, Tolerances, Distribution, Scientific research, Population.

In Section I, the moments of a normalized tolerance distribution are estimated by utilizing experimental technique deaths in the indirect assay. More precisely, the information gained by assuming that the probability of experimental technique deaths is independent of dosage may, in general, yield an LD50 with greater precision. Adjustments are given for nonconstant natural mortality over time. A preliminary report on bimodal tolerance distribution is also given. (Author)

AD-409 119 Div. 32
(TISTB/MS)

Linguistics Research Center, U. of Texas, Austin.
INTRODUCTION TO FORMATION STRUCTURES,
by D. A. Senechalle. Apr 63, 17p. Rept. no. LRC63 WTM2
Grants NSF GN54 and G19277
Unclassified report

Descriptors: (*Language, Mathematical analysis), (*Vacabulary), Theory.

Effort is directed toward a theoretical foundation for linguistic information processing algorithms which will be generally applicable to natural and artificial languages. (Author)

AD-409 120 Div. 32
(TISTB/MS)

Linguistics Research Center, U. of Texas, Austin.
THE CLASSIFICATION OF ENGLISH ADVERBIALS IN CORPUS 05,
by Howard W. Law. Apr 63, 1v. Rept. no. LRC63 WDE1
Grant NSF GN54
Unclassified report

Descriptors: (*Vocabulary, Classification), (*Language, Analysis).

Research conducted in connection with the classification of adverbials is presented in this paper. The resulting classification is tentative because, among other reasons, it deals only with data of a limited corpus. The scope of the problem and statements by some other authors are presented. The investigation involved a study of adverbial sequences and occurrences of adverbials in reference to verbals. Four classification sortings were used to aid the study. Tentative adverbial function classes were assumed. The results of the first three sortings were used to modify the tentative function classes. Tentative position classes were established. The fourth sorting was used to establish function-position classes. Criteria for de-

Figure 8b. Examples of Duplicate Indexing

promise of more objective measures of performance or quality than evaluative techniques available today.

Examples of the special factors involved in assignment indexing techniques and automatic classification include the question of the amount of computation required in the inversion and other manipulations of large matrices 1/ and the concommitant problems of how large a vocabulary of clue words can be used effectively and of whether some documents cannot be indexed at all because they contain none of these words. 2/ There is, as Needham says, "no merit in a classification program which can only be applied to a couple of hundred objects." 3/

In the various techniques for automatic clustering or categorization of documents, there are serious questions of whether the groupings can be conveniently named or displayed for the benefit of the user. 4/ Another example of special factors in the appraisal of an automatically generated classification scheme is as follows:

> "Operational testing is displeasing in that it puts off any verification until right at the end; it is expensive; there is not much experience on how to do it in a realistic way; and it is ill-controlled in the sense that the practical performance of a system is influenced by many other factors than the classification it embodies." 5/

Examples of suggested bases for evaluation made possible by machine processing itself include proposals by Doyle and Garvin, among others. Doyle in particular suggests the substitution for the elusive concept of "relevance" of criteria based on "sharpness of separation of exploratory regions in which the searcher finds documents of interest from those in which he does not find such documents." 6/ He further emphasizes the need for discriminating a particular document from other topically close documents (Doyle, 1961 [166]) and suggests that "this decision can never be made by a human---only by a computer, which is the only agency capable of having full consciousness of the contents of a library." 7/ Garvin considers the more general problems of language and meaning, and suggests that there are two kinds of "observable and operationally tractable manifestations of linguistic meaning", ---namely, translation and paraphrase, and that these may be investigated by techniques of linguistic data processing. 8/ Edmundson, however, points out that while there is in general only one translation of a document, there may be as many abstracts (and, by implication, index sets) as there are users. 9/ Thus we are back again at the questions of purpose and relevance.

1/ Compare Williams, 1963 [642], p. 162.

2/ See Maron and Borko, various references.

3/ Needham, 1963 [433], p. 8.

4/ See, for example, Doyle, 1963 [162], p. 6: "Several researchers have tried to group topically close articles, usually by statistical means, but it is rather difficult to get any benefit from this grouping unless you can represent these groups for human inspection."

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

6/ Doyle, 1963 [164], p. 200.

7/ Doyle, 1961 [169], p. 23.

8/ Garvin, 1961 [224], p. 137.

9/ Edmundson, 1962 [178], p. 4.

8. OPERATIONAL CONSIDERATIONS

Whatever the verdict of evaluation of one or more automatic indexing techniques, whether of the derivative, modified derivative, or assignment type, there are certain operational considerations and problems that typically affect any attempt to apply such techniques in actual production operations. These considerations, which also affect linguistic data processing operations in general, include input considerations, availability of methods or devices for converting text to machine-usable form, programming considerations, questions of format and content of output, and problems of customer acceptance of the machine products.

8. 1 Questions of input

Input considerations include, first, questions of the extent and availability of material which can be handled directly by the machine. This may be limited to title only, to title plus abstract, title plus other material, [1] preselected text or automatically generated extracts; or it may in a few cases extend to full running text. Possible future requirements may extend to the processing not only of full text but of interspersed graphic material (equations, charts, diagrams, drawings, photographs) as well.

We have considered typical arguments for and against the limitation of input to titles only, to augmented titles, and to abstracts in other sections of this report. The points to be emphasized here are requirements for pre-editing or post-editing, provisions for error detection and error correction, the time and cost requirements of conversion equipment if material is not already available in machine-usable form, and the like. As Cornelius suggests:

> "Present day computers, if used for machine indexing, will be generally input limited and will require excessive data preparation. Causes of these limitations are: time required for translation to machine language, verification of this machine language, and the capability or lack of capability of correction in the input media." [2]

Examples of pre-editing requirements, even for the simple case of keyword-in-title indexing, include the spelling out of chemical symbols, the encoding or the omission of subscripts and superscripts, insertions of hyphens to prevent indexing of a word, and substitutions of blanks for hyphens in compound words to assure indexing of each component. [3] For full text, a far more extensive and elaborate set of rules and conventions must be developed and applied. [4] Other editing may be required for format standard-

[1] This may specifically include cited titles, as suggested variously by Bohnert, 1962 [69], p. 19; Giuliano and Jones, 1962 [229], p. 10; Swanson, 1963 [580], p. 1; Gallagher and Toomey, 1963 [205], p. 53; and as used in the SADSACT method, see pp. 98-99 of this report.

[2] Cornelius, 1962 [140], p. 42.

[3] See, for example, Kennedy, 1961 [311], p. 120.

[4] See, for example the sophisticated proposals of Nugent, 1959 [441], and Newman et al, 1960 [439].