AUTOMATIC INDEXING

A State-of-the-Art Report

Mary Elizabeth Stevens

A state-of-the-art survey of automatic indexing systems and experiments has been conducted by the Research Information Center and Advisory Service on Information Processing, Information Technology Division, Institute for Applied Technology, National Bureau of Standards. Consideration is first given to indexes compiled by or with the aid of machines, including citation indexes. Automatic derivative indexing is exemplified by key-word-in-context (KWIC) and other word-in-context techniques. Advantages, disadvantages, and possibilities for modification and improvement are discussed. Experiments in automatic assignment indexing are summarized. Related research efforts in such areas as automatic classification and categorization, computer use of thesauri, statistical association techniques, and linguistic data processing are described. A major question is that of evaluation, particularly in view of evidence of human inter-indexer inconsistency. It is concluded that indexes based on words extracted from text are practical for many purposes today, and that automatic assignment indexing and classification experiments show promise for future progress.

1. INTRODUCTION

This report of the Research Information Center and Advisory Service on Information Processing (RICASIP) [1/] is one of a series intended as contributions to improved cooperation in the fields of information selection systems development, information retrieval research and mechanized translation. In each of these areas, automatic techniques for linguistic data processing are receiving increased attention. This report covers a state-of-the-art survey of current progress in linguistic data processing as related to the possibilities of automatic mechanized indexing. Insofar as has been practical, the survey of the literature on which this report is based has been made through February 1964.

It has concentrated on the major developments in and related demonstrations of automatic indexing potentialities. Examples are also given of indexes compiled by machine and of potentially related research efforts in such areas as natural language text searching, statistical association techniques used for search and retrieval, and proposed systems for concept processing. There are, undoubtedly, various omissions. Neither the inclusion of reports on various specific experiments and techniques nor the omission of others is intended to reflect an endorsement as such of those that are included or an adverse evaluation of those that are not mentioned.

1/ Initiated at the instigation of the National Science Foundation. RICASIP is jointly supported by NSF and NBS.

1.1 Definitions and Background

The noun "index" has as its most general meaning "something used or serving to point out, a sign, token, or indication", (American College Dictionary) or "that which shows, indicates, manifests, or discloses; a token or indication" (Webster's International Dictionary, 2nd Edition, unabridged). More specifically, an index is "a pointer or key which directs the searcher to recorded information." [1/] The terms "index" and "indexing" have been used in the fields of library science and documentation with reference to the fact that the selection of information pertinent to a particular problem or interest, from all the previously recorded information available, involves problems of decision-making based on less than the full content or text of each of the records being searched.

Short of complete scanning of all the possibly relevant material, it is necessary to select or "distill" condensed representations or surrogates [2/] for each item. These surrogates are intended to direct the searcher to the most probably pertinent items in a collection. The operations known as "indexing" thus involve:

(1) Choosing clues that will serve to identify, for purposes of later retrieval, a particular book, document, or other recorded item, and

(2) Either marking on the item itself or recording as a separate item-surrogate the tags, labels, or codes representing these clues.

The second of these two steps can be purely clerical in nature, but the first has been, to date, primarily the result of human intellectual efforts in subject content analysis.

Well-known inadequacies of human indexing operations include both those stemming from man himself and those which result from the volume and the character of the materials with which he deals. On the human side, there are fundamental questions of perception, comprehension and judgment, as well as those of inter-indexer and even intra-indexer consistency. In addition, the indexer is asked to guess in advance what others will ask for, understand, and find relevant on future search. He is even asked, in effect, to anticipate the language of future inquiries. Thus, a somewhat facetious definition of the noun "index" has a considerable sting of truth: "A system of analyzing information in which the method used to choose categories is carefully hidden from the user. An attempt to outguess the future." [3/]

The nature of the material to be indexed, especially in the area of scientific information, raises a number of crucial problems. The still increasing spate of production of technical literature and reports poses not only the problems of sheer volume in terms of

---

1/ Crane and Bernier, 1958 [144], p. 513.
(Note: Full citations of references are given in the bibliography by author and by numerical order of the figures in brackets.)

2/ See, for example, R. E. Wyllys, 1962 [651], for discussion of the two-fold purposes of condensed representations: to serve a search-tool function on the one hand and a content-revealing one on the other.

3/ Vanby, 1963 [622], p. 143.

manpower requirements and time necessary to produce indexes, but also problems of glut in terms of man-hours necessary for the individual scientist to maintain awareness of what is going on in his field. There are major problems created by newly emerging fields of effort, new interdisciplinary areas of interest, and dynamically evolving terminology. Increasing specialization, on the other hand, brings out additional difficulties in finding what has been done elsewhere that might be applicable to one's own work and in avoiding wasteful duplication of effort, with their own attendant problems of terminology.

All these problems are aggravated by the increasingly critical urgency which should apply to making all useful information available to those who need it as promptly and as selectively as possible. Recognition of this urgency and of the inadequacies of present solutions has therefore prompted consideration of the feasibility of using machines to assist in the indexing process.

The term "mechanized indexing" signifies the accomplishment of some or all of the indexing operations by mechanized means. The term includes the use of machines to prepare and compile indexes, and to sort, assemble, duplicate and interfile catalog cards carrying index entries. In this report, however, we shall be concerned primarily with the area of automatic indexing, that is, the use of machines to extract or assign index terms without human intervention once programs or procedural rules have been established. This term is chosen in preference to auto-indexing as originally suggested by Luhn (1961 [373]) for the reasons set forth by Bar-Hillel, [1/] and to machine indexing [2/] due to possible confusion with machine tool operations. Automatic indexing has been used by such workers in the field as Gardin (1963 [209]), Kennedy (1962 [310]), Maron (1961 [395]), Swanson (1962 [584]), and Wyllys (1963 [653]).

For obvious reasons, we also subsume under this term any specifically "clerical" (Fairthorne, 1956 [188], 1956 [189], 1961 [190] and hence machinable operations that can similarly be substituted for human intellectual effort. There is nothing that machines can do which people cannot do except for limitations of time, cost, or availability of appropriate resources. Thus, we shall consider "machine-like indexing by people" (O'Connor, 1961 [447]; Montgomery and Swanson, 1962 [421]) as falling properly within the scope of automatic indexing, especially in the sense of "... deciding in a mechanical way to which category (subject or field of knowledge) a given document belongs ... deciding automatically what a given document is 'about'." [3/]

The principle of indexing, that is, of using subject-content clues and item surrogates as substitutes for searches based on perusal of the full contents, has a history of several millenia. In ancient Sumaria and Babylon, clay tablets were sometimes covered with a thin clay envelope or sheath that was inscribed with brief descriptions of the contents of the tablet itself (Carlson, 1963 [101]; Hessel, 1955 [268]; Lalley 1962 [343]; Olney, 1963 [458]; Schullian, 1960 [525]). The first known instance of an index list is apparently that of Callimachus in the third century B. C., which was a guide to the contents of some 130,000 papyrus rolls (Olney, 1963 [458]; Parsons, 1952[469]).

1/ Bar-Hillel, 1962 [35], p. 417.

2/ Bohnert, 1962 [69]; Edmundson, 1959 [176]; and others.

3/ Maron, 1961 [395], p. 404.

Application of the indexing principle by use of clerical procedures that today can be accomplished by machine was suggested a little more than a century ago. A British librarian, Andreas Crestadoro, advocated the permutation of the words in titles in 1856, claiming that thus the subject matter index would follow the author's own definition of the contents of his book. He prepared such "concordances of titles" for several different library collections. 1/

Within a generation, punched card machines had been invented, but they were not to be used for library and documentation purposes for some decades yet. 2/ Keppel, writing in 1937 of his vision of the library 21 years in the future, says:

> "When it comes to using the cards, I blush to think for how many years we watched the so-called business machines juggle with payrolls and bank books before it occurred to us that they might be adapted to dealing with library cards with equal dexterity. Indexing has become an entirely new art. The modern index is no longer bound up in the volume, but remains on cards, and the modern version of the Hollerith machine will sort out and photograph anything the dial tells it ..." 3/

By 1945, Bush had prophesied Memex [93], and in the 1950 Windsor lectures Ridenour referred to an RCA development, the so-called "electronic pencil", a proposed reading aid for the blind intended to convert printed characters to a suitable coded form. He went on to suggest:

> "... We shall have to arrange for cataloguing to be done by machine, without human interaction except in terms of setting up once for all the system on which the cataloguing is performed... It is only a step from this device (the electronic pencil) to the electronic catalogue, which will read text for itself, recognize key symbols and phrases with which it has been provided, and construct appropriate catalog entries for the text it reads." 4/

It has only been in the past decade or so, however, that there have been any serious efforts directed to the use of machines for automatic indexing. In the period 1957-1958, Luhn first presented and published several provocative papers dealing with such challenging possibilities as "auto-abstracting", "auto-encoding" and "auto-indexing" (Luhn, 1957 [385]; 1958 [374]; 1959 [371]). Luhn's work on the permutation of significant words in titles, abstracts, and complete text, the Keyword-in-Context or KWIC

1/ See Crestadoro, 1856 [146]; see also Farley, 1963 [192]; Linder, 1960 [362]; Metcalfe, 1957 [416]; and Ohlman, 1960 [451].

2/ See pp.19-22 of this report.

3/ See Keppel, 1939 [316], p. 5.

4/ See Ridenour, 1951 [500], p. 26.

system, also began about this time. [1] Also in 1958, Baxendale published the results of experiments in automatic indexing involving scanning of topic sentences, syntactical deletion processes and automatic phrase selection (Baxendale, 1958 [41]).

With respect to the KWIC and permuted title techniques, several independent approaches were being developed at about the same time as Luhn's. These concurrent efforts were carried out at the Wright Air Development Center (Netherwood, 1958 [437]), the Rocketdyne Division of North American Aviation (Carlsen, et al, 1958 [99]), and the System Development Corporation (Citron, et al 1958 [120]; Ohlman, 1960 [451]). [2] Netherwood's permuted title index to a bibliography on logical machine design involves manual simulation of a machineable method. Although the results were not published until June 1958, the manuscript was submitted in November 1957. [3] The Rocketdyne permuted-title bibliography, on industrial control, is credited by both Henderson (1962 [263]) and Ohlman (1960 [451]) as the first to be produced on computers, the program

---

1/ In a private communication dated March 13, 1963, Luhn provided the following chronology:

May 1957	Routine 1 Program for word isolation within 60 characters per card, written by H. C. Fallon.
1957-1958	Creation of concordances of various scientific papers in the form of cards, each card showing a keyword centrally located within 60 letters worth of the associated phrase. Experimentation with these cards to arrive at thesauri for special fields of interest or study. Idea of automatic indexing by means of significant or keywords in context conceived by H. P. Luhn.
May 1958	Keyword-in-Context Index for titles only initiated by H. P. Luhn and samples produced with Routine 1 Program.
June 1958	Start punching of titles for Keyword-in-Context Index for literature on Information Retrieval and Machine Translation. (Keypunching done by Miss Olive Ferguson.)
August 1958	Simplified version of Routine 1 written by H. C. Fallon for generating Keywords-in-Context Indexes and delivered to Service Bureau Corporation, New York City.
September 1958	First Fdition of Bibliography and Keyword-in-Context Index on Information Retrieval and Machine Translation published by Service Bureau Corporation.
January 1959	Started writing program for improved version of Keyword-in-Context Index, including derived identification code, written by Jr. J. Havender.
June 1959	Second Edition of Bibliography and Keyword-in-Context Index on Information Retrieval and Machine Translation, published by Service Bureau Corporation, including derived identification codes.

2/ See also National Science Foundation's CR&D Report No. 3, [430], p. 39.

3/ Netherwood, 1958 [437] , p. 155, footnote.

having been written by J. T. Madigan. [1] At any rate, both this program and Luhn's KWIC program at IBM were apparently written relatively early in 1958.

Citron et al (1958 [120]) in presenting results of the SDC work and Ohlman in his chronological bibliography of permutation indexing (1960 [451])cite as at least partial predecessors the "rotated file" principles developed at the Chemical-Biological Coordination Center (1954 [112]; Heumann and Dale, 1957 [270] and 1957 [271]; Wood, 1956 [649]). It should also be noted as a matter of historical background that a system for machine manipulation and compilation of permuted title-and-term-index records has been in productive operation since 1952. [2] This earlier effort was not generally known to other investigators and was apparently first reported in the open literature as late as 1961.

Notwithstanding such other efforts, it is conceded by almost all workers in the fields of automatic abstracting and indexing that the major credit for pioneering interest and impetus should be attributed to Luhn and Baxendale. Specific acknowledgements of their "pioneering work" and "first steps" have been made by many investigators both in this country and abroad--for example, Borko and Bernick, [3] Hines, [4] Mooers, [5] Pevzner and Styazhkin, [6] and Wyllys. [7] In particular, the Russian investigator Purto states: "So far as we know H. P. Luhn was the first investigator to suggest the concept of a set of significant words for the consideration of problems in automatic abstracting." [8]

Much of the early effort 1957-58, whether at IBM or elsewhere, was in fact spurred on by the International Conference on Scientific Information (ICSI) held in Washington, D. C., in November, 1958. The printed text of both the Preprints [478] and the final Proceedings [480, 481] was deliberately prepared, over the typographer's objections, so that a double space followed each period ending a sentence, in order to facilitate machine processing of this text. Thus the printers ".... were faced with ... the necessity to prepare the final volume of the Proceedings from these preprints, and to arrange type composition amenable to computer analysis. The latter is an experiment. With an eye to the distant future, the Program Committee wished to make available the monotype punched tapes from the text for statistical studies with computers. We hope

---

1/ Carlsen, et al, "Information Control", 1958 [99], p. 20.

2/ Veilleux, 1962 [624], p. 81: "Consumer demand balanced against availability of manpower and machine time were the factors which led to the establishment of the permutation title word indexing project in 1952."

3/ Borko and Bernick, 1962 [77], p. 3.

4/ Hines, 1963 [273], p 7.

5/ Mooers, 1963 [424], p. 4.

6/ Pevzner and Styazhkin, 1961 [472], p. 3.

7/ Wyllys, 1961 [650], pp. 6-7.

8/ Purto, 1962 [484], p. 2.

some work of this kind will be demonstrated during the Conference. This has caused some compromises in typography..."[1/]

Several pioneering experiments in automatic indexing were applied to this ICSI material. One of these led to the preparation of a permuted keyword index based on titles, subtitles, section and table headings, figure captions, and selected sentences or phrases taken directly from the text (Citron, et al, 1958 [120]). It was prepared using punched card equipment, and the resulting listings were distributed to the Conference participants in November of 1958. Another set of experiments involved trial of the "auto-abstracting" and "auto-encoding" techniques proposed by Luhn (1958 [379]).[2/] A computer program potentially applicable to certain ancillary operations which might be involved in automatic indexing was also demonstrated at the time of the ICSI sessions. (Stevens, 1959 [568]).

Much of the rapidly proliferating work in the field of automatic indexing since that time has been inspired directly or indirectly by the results of these experiments using the ICSI material. For example, Dowell and Marshall, discussing early efforts at the English Electric Company, state: "We first became interested in the possibilities of computer produced indexes through Luhn's work at IBM and the early examples of KWIC indexes which were distributed at the time of the Washington Conference..." (Dowell and Marshall, 1962 [159]).[3/]

1/ "Preprints of papers of the International Conference on Scientific Information," 1958, [478], Preface. (The monotype tapes are in fact still held in the custody of the Research Information Center and Advisory Service on Information Processing, National Bureau of Standards, but difficulties to be discussed later in this report discourage their use.)

2/ See also his "Automated intelligence systems" 1962 [372], note 11, p. 100: "Papers for this conference were distributed to participants two months ahead for study. By arrangement with the Columbia University Press the Monotype tapes used in publishing these preprints were made available for experimentation. At the conference exhibit, IBM researchers demonstrated the automatic transcription of these Monotype tapes to magnetic tape via punched cards and thence the automatic creation and printout of abstracts by means of electronic data processing equipment at the Space Systems Center in Washington, D. C. All this was done without any human intervention except for the handling of the input and output records. Also, preprinted Auto-Abstracts of Papers of Area 5 of the Conference were made available to participants at the beginning of the conference."

3/ See also R. A. Kennedy, 1962 [310], p. 181: "While automatic indexing in any interpretative and analytical sense is therefore not yet a practical matter, a simpler mode of machine indexing is coming into wide use ... primarily stimulated by the publication in 1958 and 1959 of reports by Ohlman, Hart and Citron and Luhn."

A somewhat premature attempt was made to establish a subscription service for KWIC indexes for a number of journals, for initial distribution beginning January 1, 1959. [1] Called PILOT (Permutation Indexed Literature Of Technology), the proposed service was advertized as "a revolutionary new totally cross-referenced index ... and it will be produced at the speed of light". Figure 1 is a reproduction of a part of the brochure issued in 1958 by Permutation Indexing, Incorporated, Sol Grossman, President, Los Angeles. While, perhaps unfortunately, the number of subscription orders received was not adequate in terms of the ambitious coverage planned, work on permuted title indexing elsewhere did lead rapidly to the publication of such indexes on a production basis.

As of February 1964, there are more than 40 examples of KWIC and other variations of permuted keyword indexing techniques in productive operation or available to the searcher. KWIC-type techniques have also been extended to special one-time index compilations and other applications, as in "automated content analysis" of verbal protocols of psychiatric interviews and group leadership training sessions (Ford, 1963 [198]; Hart and Bach, 1959 [256]; Jaffe 1962 [294] and 1958 [296]; Stone, et al, 1962 [575]).

The same period during which the ICSI was planned and held (1957-1958) was also marked by the first issue of Current Research and Development in Scientific Documentation by the National Science Foundation. In it and in subsequent issues, there were reported other early efforts in machine-compiled indexes, in the construction and use of special thesauri, and in indexing and retrieval experiments based on machine processing of text. Thus, for example, punched card methods for compiling printed indexes and announcement lists were under consideration at Bell Laboratories and at Esso Research and Engineering. Special attention was being given to thesauri as early as July 1957 at both Chemical Abstracts Service and the Cambridge Language Research Unit, and at Ramo Wooldridge, "Research on the problems of fully automatic indexing and retrieval based on raw text input to a general-purpose computer is under way." [2]

Nevertheless, as of the present date, the question of the possibility of automatic indexing in the sense of the substitution of machineable procedures for human intellectual efforts normally required to identify, categorize, classify, index, select, and list particular items in a collection of items is still moot. Opinions run the gamut from extreme pessimism, "Mechanization of abstracting and indexing is rejected as impractical for the foreseeable future" [3] to enthusiastic optimism, "The conclusion that automatic indexing and cataloging is superior to human indexing and cataloging is both provocative and remarkable." [4]

Borko and Bernick claim that " ... Raw data, i.e., unedited natural language text, can be processed statistically so as to automatically assign index terms to each document and to classify the document into a subject category; this has been demonstrated." [5] On the other hand, Farradane thinks that any form of mechanized processing in indexing

---

1/ See Linder, 1960 [363], p. 99 and Figure 1.

2/ National Science Foundation's CR&D Reports No. 1, [430] pp. 4, 6; No. 3 [430], pp. 12, 19, 31.

3/ Bar-Hillel, 1958 [33], abstract.

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

5/ Borko and Bernick, 1963 [78], p. 28.

Figure 1. Brochure for Proposed Permuted Title Index Service

operations is "liable to continuous error", [1] while Baxendale takes a middle ground: "Thus far the role of the computer is chiefly that of research instrument; whether or not it can fully assume the task of indexing is still in doubt". [2]

1.2 Scope of This Study

In view of the continuing controversy over the feasibility and evaluation of automatic indexing techniques, a state-of-the-art survey and report is perhaps premature at this time. The topic is controversial on at least five grounds: First is the question, "Can indexing be done by machine at all?" Next,"Is what can be done by machine properly termed 'abstracting', 'indexing', or 'classifying'?" The third moot point is "Is whatever can be done by machine good enough, acceptable, as good as, or better than the product of human operations?" The fourth and most critical question is "How can we evaluate acceptability or comparability for any indexing process whatsoever, whether carried out by man or by machine or by machine-aided manual operations?" Finally, "If an indexing product is to be achieved by machine, can it be done by statistical means alone, or must syntactic, semantic and pragmatic considerations be brought to bear in the machine decision-making processes?"

The heat of controversy over any of these five grounds of debate is almost inversely related to the availability of objectively validated evidence to which appeal might be made. Thus, the literature on the topic to date is typically colored by personal reactions both pro and con, and even the cynics rely more on subjective judgments and personal preferences than on any substantial body of data. O'Connor cites typical claims of both proponents and opponents of the feasibility of automatic indexing, and he comments on both, "I have seen no good evidence offered in support of such a conclusion." [3]

An impartial middle ground is offered by recognition that "To define a process ordinarily thought to require human intellectual effort in such a way that it can be performed by a machine imposes a rigor and a discipline on the definition which itself is invaluable to understanding the nature of the process". [4] Learning more about the indexing process itself, through experimentation with machines, will provide "results of general interest, not just to those optimistic about machine indexing experiments". [5] In this sense, a state-of-the-art study is not premature. In this sense, therefore, we shall explore the five questions listed above in subsequent sections of this report.

1/ Farradane, 1961, [193], p. 236.

2/ Baxendale, 1962 [42], p. 69.

3/ O'Connor, 1961 [447], pp.274 and 275.

4/ Swanson, 1962 [583], p. 288.

5/ Bohnert, 1962 [69], p. 9.

More particularly, in this survey of automatic indexing efforts, we will be concerned with the following principal topics:

(1) A brief indication of the variety of ways in which punched card machines and computers can be and have been used in the preparation or compilation of indexes. 1/

(2) A more detailed consideration of the possibilities for machine generation of indexes, specifically including:

  (a) Automatic derivative indexing, as in various examples of machine extraction of keywords, where selection is based upon pre-specified criteria,

  (b) Automatic assignment indexing, whereby the machine is programmed to determine, in accordance with various specified criteria, whether or not some one or more members of an established list of 'labels' (such as subject headings, class names, descriptors, or other indexing terms) should appropriately be assigned to the document or item in question, and

  (c) Automatic classification techniques, on which such assignment-indexing operations may or may not be based.

(3) Consideration of the use of machines as relatively sophisticated aids to human intellectual operations applied in either subject-content analyses or search-strategy determinations.

(4) Discussion of the question of evaluation of any index whatever, whether manually or mechanically prepared.

(5) Consideration of the implications of related research and development efforts, specifically including:

  (a) Comparative evaluation of indexing systems,

  (b) Development and use of new types of "indexing" aids (in the sense of "pointing to" and "indicative of" the probable subject-content relevance) to either selective dissemination or retrospective search of the technical literature,

  (c) Linguistic and logical-inference approaches to the elucidation of 'meaning' in natural-language messages, and

  (d) Theoretical approaches to the problems of determining "membership-in-classes".

1/ Note that card-controlled camera systems, such as the Listomatic, and Addressograph machines have also been used for index compilations. See, for example, Shaw, 1951 [542], p. 49, who cites early use of the Addressograph for bibliographical work by A. Predeek, "Die Adrema-Maschine als Organizationsmittel im Bibliotheksbetriebe", Berlin, 1930, and E. Morel, "Les Machines au secours de la Bibliographie", Revue du Livre 1:14-19 (1933). Use of such devices is not included in this report, however, since they cannot be adapted to machine generation of indexes.

(6) Appraisal of the current prospects for further research and development.

Certain difficulties of organization are evident. Thus many proposals precede actual tests of techniques to which they are akin. Other proposals have been engendered as by-products of or incidental to investigations of other techniques, such as those of text processing to derive by machine selected sentences which together may serve as automatically generated "abstracts", more properly extracts. 1/

This related subject of automatic abstracting, i.e., the application of machine-usable rules to the extraction or generation of textual information representing in condensed form that carried in the document as a whole, will not be of primary concern. However, it will be noted that most of the automatic abstracting techniques so far proposed are potentially usable as tools for automatic indexing, especially in the trivial sense that the automatic selection of index terms could be based solely upon the substantive words found in the machine-prepared extract. 2/ Further, since we are presuming that a state-of-the-art review of automatic indexing techniques is in some sense appropriate at this time, we shall emphasize the actual results of machine compilation and machine generation of indexes and those investigations of assignment-indexing techniques for which experimental or comparative data have been reported, rather than theoretical approaches.

1/ See, for example, Luhn, 1959 [384], p. 4: "The principle of abstracting information by extracting certain portions or elements from the full text of a document is particularly suitable to mechanization"; Becker, 1960 [44], p. 13: "Perhaps 'extracting' would have been a better word than 'abstracting'"; Edmundson and Wyllys, 1961, [181], p. 227: "All proposed methods for making an automatic abstract of a document involve using the author's own words by selecting complete sentences, thereby reducing abstraction to the simple task of extraction."

2/ See Wyllys, 1963 [653], p. 22: "Automatic indexing is an area that seems to us to be especially close to automatic abstracting, since the words and word groups found to be most representative of a document for automatic abstracting purposes are obvious candidates for entries in an automatic index for the documents." See also Tanimoto, 1961 [594], p. 235: "Thus after extracting k sentences which are a predetermined small fraction of the document, we have an 'abstract'. To find the indexes to the document we take these k sentences and the corresponding sets of the canonical elements and consider terms versus sentences instead of sentences versus terms... The same analysis is then applied to this 'transposed' problem to produce the index terms"; Yakushin, 1963 [654], p. 17: "If some method can be employed for the automatic compilation of abstracts, it can as well be used for the subject index."

1.3 Derivative vs. Assignment Indexing

At least part of the provocation and controversy with respect to the possibilities for the use of machines in indexing is due to confusion as to what type of indexing is meant. This in turn relates to a much older and broader controversy--that between "word" or "catchword" indexing on the one hand and "subject indexing", "concept indexing", or "controlled indexing" on the other.

In terms of operational definition, the contrast is best expressed in Luhn's distinction between index entries that are derived from the text of an item itself and those that are assigned to it from a list or schedule of subject categories, descriptors and the like, which exists independently of the text of the item (Luhn, 1962 [372]). 1/ In general, the differentiations that are made for the broader controversy, and the claims and counter-claims made by the enthusiasts of either school, provide background for the distinctions that should be made between various automatic derivative indexing operations and whatever possibilities may be demonstrated for assignment indexing by machine.

In his text on information storage and retrieval Kent (1962 [315]) contrasts word indexing as used in permuted keyword indexes, concordances and "pure" uniterm systems with controlled indexing which "implies a careful selection of terminology used in indexes in order to avoid, as far as possible, the scattering of related subjects under different headings." He notes elsewhere that word indexing requires little subject-matter training on the part of the indexer and little skill in indexing as such, and adds: "It is this type of indexing that a machine can perform well." 2/

Like Kent, Bernier thinks that true subject or assignment indexing requires highly trained human indexers. He says further:

> "The difference between subject and word indexing has been unclear at times. Both types employ words, but only true subject indexing employs them with discrimination. Word indexing leads to omission of entries, scattering of related information, and a flood of unnecessary entries. Word indexing uses words as they are found in the material indexed with a minimum regard for standardized meaning..." 3/

Herner provides a further amplification of differences that are pertinent to consideration of indexing by machine, as follows: 4/

1/ See also Herner, 1962 [266], p. 5; Skaggs and Spangler, 1963, [557], p. 60; Slamecka, 1963 [558], p. 224. Mooers makes a similar distinction between "index terms which are words or phrases extracted from the text and stylized conceptual terms--cliches --which are assigned to the text", 1963 [423], p. 4.

2/ Kent, 1962 [314], p. 268.

3/ Bernier, 1956 [54], p. 23.

4/ Herner, 1963 [267], p. 183.

"The differentiation that is made between the two types of indexing is that word indexing is inextricably tied to the words in a text: If a word appears it gets indexed as such; if it does not appear it does not get indexed. Concept indexing, on the other hand, has an element of abstraction in it: Words may either be indexed as such or may be converted, either by themselves or in combination with other words, into concepts which may not bear a direct resemblance to the words or combinations of words that evoked them in the indexer's mind."

Machine techniques such as those of Luhn's KWIC, like the early Uniterm systems, look no farther than the words used by the one author himself. Techniques such as those of Maron, Swanson, Borko, Meadow and Williams, among others, look specifically to relationships between words as used by one author to patterns of word usages in a given subject area or given document collection. They may also look to these patterns as in turn related to prior human analytic judgments of the "aboutness" referrents of items in the collection. In this sense, they at least attempt replication by machine of assignment indexing.

There is no real question but that machines can in fact derive words from text provided that it is in machine-readable form. This machine procedure may involve direct extraction of all words as index entries, as in a complete concordance. It may involve the extraction of only those words which survive a "purging" operation in which articles, conjunctions, adjectives, and other "common" words are first deleted. Various machine-controlled modifications to such "derivative" indexing are also available. The case for machine achievement of assignment indexing for any but limited special cases is not so clear.

2. INDEXES COMPILED BY MACHINE

A first and obvious use of machines in indexing processes is in the manipulation of index entries, previously selected on the basis of human analysis, to produce various orderings, duplications and listings of these entries. The power of machine techniques to speed and economize the sorting, ordering and listing operations in the preparation or compilation of indexes was recognized quite early, both in the field of library science and in the consideration of potential areas of application by specialists in machine potentialities.

In particular, two specialized types of index, at least in the broad sense, are such that their compilation would be almost prohibitive in terms of time and cost were it not for the use of machines. These are, respectively, the case of the complete index, the index to all words of a text in their various contexts, which is a concordance, 1/ and the case of the "citation index", which has been used in the field of law for many years but has only quite recently been suggested for literature search purposes related to scientific and technical information.

1/ See, for example, Doyle, 1963 [162], p. 11: "Without data-processing machinery, concordances are prohibitively expensive to generate for most uses except in those cases where it is well known that a given volume of text is going to be used again and again, by large numbers of people over a long period of time. As we know, clergymen have made use of manually prepared concordances of the Bible since the 12th century".