POTENTIALITIES OF AUTO-ENCODING OF SCIENTIFIC LITERATURE

by

H. P. Luhn

International Business Machines Corporation
Research Center
Yorktown Heights, New York

ABSTRACT: The introduction of mechanical devices for the processing of scientific information raises the question as to the extent to which machines will be able to assist in the selection, storage, dissemination and retrieval of information. In order to appreciate fully the functions that information processing machines are capable of performing in this area a number of typical operations are presented and their potential usefulness to the development phase as well as operational phases of information systems is explored. The solution of particular problems is illustrated by way of examples based on the availability of scientific literature in machine-readable form. The examples cover the compilation of word lists, establishment of word relationships, the preparation of word patterns for retrieval and the compilation of dictionaries and thesauri. Some of the results of Information Retrieval Research at the IBM Research Center are presented in the form of machine print-outs such as the keyword-in-context index for bibliographies, the auto-abstract, the word pair matrix, derived code words, and the statistical analysis of a document.

Research Report
RC-101
May 15, 1959

Printed in U. S. A.

Table of Contents

POTENTIALITIES OF AUTO-ENCODING OF SCIENTIFIC LITERATURE

By H. P. Luhn

Introduction

The introduction of mechanical devices for the processing of scientific information in the form of documents raises the question as to the extent to which machines will be able to assist in the selection, storage, dissemination and retrieval of information. Information of this kind is generated not as a continuous stream but in self-contained parcels dealing with but a small fraction of the total universe of scientific knowledge. The normal motivation for producing such parcels or 'documents' is to relate new information i. e. something which is different, in certain respects, from anything that has been reported previously. Yet, whatever it is that is different can only be expressed by relating it to things which are known at the time because of previous communications. There is, therefore, a certain degree of similarity between a new document and certain others which preceded it. It is this overlap of communicated knowledge which forms the basis for characterizing and organizing new information for the purposes set forth above.

The extent of similarity of a new document with past documents and the extent of its novelty can be judged by humans and be expressed by means of comparatively few classifying terms. Presently a great deal of research and development effort is directed at the discovery of automatic methods by which documents may be characterized by appropriate terms which, in turn, may serve to establish similarity between documents. The potentialities of such 'auto-encoding' methods and of certain associated procedures will be explored in the following.

Availability of Machine-Readable Information

Automatic processing of information demands that the text of documents be available in machine-readable form, such as punched cards, punched or magnetic tape. At which stage in the process of document creation the transformation to a machine record may take place depends on circumstances, procedures, and the availability of appropriate devices. The earlier this transformation can be accomplished the greater are the savings that may be realized not just in deriving machine records for the purpose of information processing but also in the overall process of producing the finished document.

Looking at the various stages in reverse order, the transcribing of the finished text of a published document involves the greatest effort. Of course, in many instances there is no choice as in the case of existing documents or where the creation of documents is beyond control or influence of those interested in the machinable version of such documents. Under these circumstances the documents will have to be transcribed by hand or eventually by print reading devices. In either of these cases error-free copy may not be realized without proofreading. The expense of transcribing such documents in their entirety will be justifiable to a limited extent only and it may, therefore, be assumed that automatic processing will be mainly applied to future literature.

A next earlier instant where machinable transcripts of texts may be obtained is at the stage of typesetting, provided a document is to be printed from type. Many typesetting processes are performed in two steps, the first consisting of creating a tape by means of a keyboard-operated punching device and the second consisting of feeding this tape into an automatic type casting or photo typesetting machine. By making this tape available to the information processor, suitable transcripts into the machine language of the processing equipment may be created by fully automatic means with little effort. The machine records obtained by this method contain errors and it depends on subsequent applications of these records as to whether such errors may be tolerated. If not, the process of proofreading needed to obtain error-free copy of the printed document will have to be duplicated for the machine record, adding an appreciable amount of effort to this process.

Certain typesetting machines are operated directly, without the intervention of punched tape. The keyboards of these machines may readily be modified to permit the creation of machinable tape as a by-product.

A still earlier stage at which machinable records may be obtained is at the moment when typewritten text is produced. Typewriters that create punched cards or tapes while typing are commercially available. However, their use will probably be limited to large organizations where many reports are produced and where the processing of these reports in machinable form is desired. The problem of error correction with these machines is greatly simplified. The use of tape-producing typewriters may be eventually extended to eliminate the retyping of text for the purpose of typesetting. In this case a fully corrected tape may be created which may serve as the input to both the typesetting machine and information processing machine, thereby eliminating the duplication of proofreading otherwise required.

Automatic Processing of Documents

Once documents are available in machine readable form in their entirety, a decisive step will have been made toward the automation of document selection, storage, dissemination and retrieval. The availability of full text offers complete freedom as to the particular methods which may be developed to accomplish the functions just enumerated. However, this freedom does not by itself guarantee an ultimate perfect solution of the many problëms involved.

The process of discovering similarity amongst documents or parts thereof needs to be carried out on various levels, depending on the particular function an information system is to fulfill. There exists, therefore, the problem of adjusting the extent of processing in such a manner that the specific requirements may be satisfied with a reasonable degree of efficiency. The tendency has been to condense the information contained in a document into a statement which characterizes the document to the extent required by the system. Statements of this kind are typified by titles, abstracts and index entries on the one hand and by class assignments, subject headings and keywords on the other. The intellectual processes performed in all these cases are abstracting processes in the broad sense.

The principle of abstracting information by extracting certain portions or elements from the full text of a document is particularly suitable to mechanization. The problem is to determine which portions or elements are most appropriate for a given situation and what operation will have to be performed with them to derive the kind of characteristics needed.

In order to appreciate fully the functions that Information Processing Machines are capable of performing in this area, a number of typical operations will be presented here. Their potential usefulness applies to the development phase as well as the operational phases of information handling systems.

Compilation of Word Lists

This operation is basic in connection with the development of systems. The vocabulary for a given discipline is necessarily the foundation for the establishment of useful criteria by which documents relating to this discipline may be identified or classified. If keywords are to serve this purpose in the retrieval operations of a system, for instance, there must be a way of telling whether a given keyword is actually contained in any of the documents of the collection to be searched.

Because of the appearance of new words, as time goes on, it is necessary to update such lists. Therefore, it would become a standard operation to extract a word list from each new document entering a system. This list would at the same time serve to establish retrieval patterns for that document.

Word lists are derived from a document by feeding the complete text into an information processing device. The individual words are sorted in alphabetic order. In conjunction with this process certain data may be ascertained and certain transformations be performed. Thus the number of occurrences of a word may be registered and listed with each word, or the location of each word within the document may be appended. Words of the same stem but of different endings may be consolidated into one standard word form. Certain words may be omitted from such lists in accordance with a list of exceptions stored in the machine memory.

The individual lists may be merged into a combined list covering a document collection of a specific area, thereby furnishing a vocabulary typical of this area.

In all these cases the lists may alternatively be given in the order of frequency of occurrence of the words so as to facilitate the recognition of word ranking and the selection of useful keywords in accordance with such ranking.

Establishing Word Relationships

This operation promises to be useful in certain applications where the characterization of documents by means of isolated words fails to bring about the desired degree of discrimination. If certain words could be given in their relationship to other words, more specific meanings may be identified by such combinations. These relationships may range from the mere co-occurrence of certain words within a phrase or sentence to the combinations of specific parts of speech.

Information Processing Machines may be programmed to carry out complex operations on the text of documents for the purpose of selecting and extracting certain portions of word combinations from the text.

Preparation of Indexes

A rather elementary operation is the selection of portions of sentences which have certain words as their nucleus. These words may be keywords from a list previously established on the basis of frequency of occurrence in the document or other criteria. A certain number of words preceding and succeeding such keywords could be lifted from the text together with the keyword itself and be presented as a means for amplifying the meaning of this word. This procedure is basic for the creation of concordances. * It may also serve to create indexes in a purely mechanical manner.

A first degree of such "Keywords-in-Context Indexes" may be derived from just the titles of a collection of documents. By having the keywords assume a fixed position within the extracted portion and by arranging these portions in alphabetic order of the keywords, a bibliographical index may be compiled. If there are several keywords in a title, as many such listings would be given as there are keywords. The format of the keyword-in-context index as applied to document titles is illustrated by way of a sample page, Fig. 1.

A more informative auto-index might include the extraction of index entries from the abstract of the document or even from the complete text. This latter procedure would lend itself to the compilation of indexes for books even if only to the extent of furnishing the indexer with a complete listing for his analysis and selection.

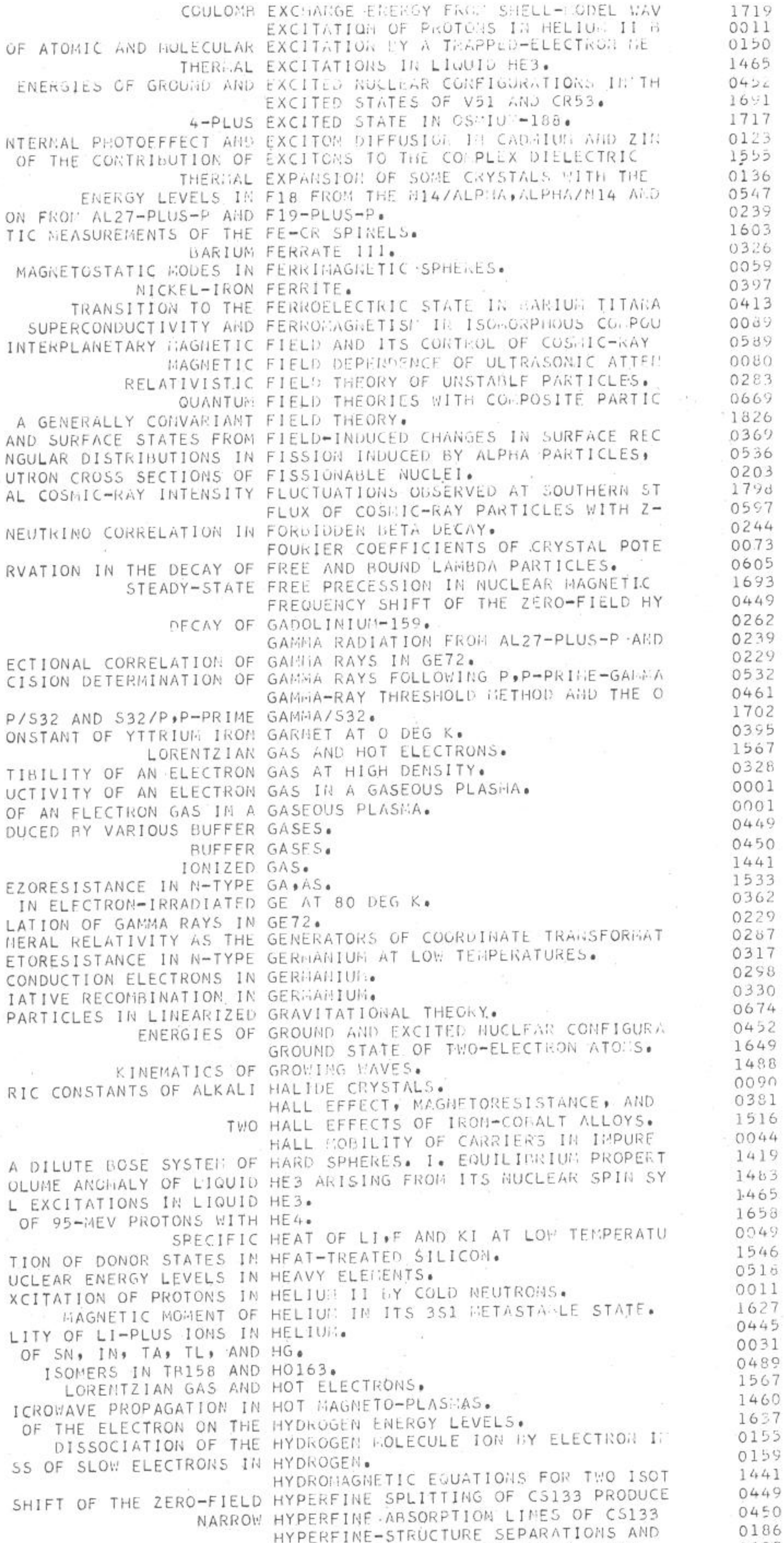
Preparation of Auto-Abstracts

A more complex process that may be performed by machines is that of selecting whole sentences from the text of a document. These sentences may be chosen not only according to the presence therein of certain words but also with respect to the relationship of these words to each other in terms of physical location within a sentence. For instance it may be argued that if a sentence contains more of certain high frequency words in closer proximity to each other than other sentences, that such a sentence is more representative of the subject matter discussed, than other sentences. If this argument is valid, then a statistical method will have been found for producing abstracts of documents by automatic

* P. Tasman, "Literary Data Processing", IBM Journal of Research and Development, July 1957.

Fig. 1

KEYWORD-IN-CONTEXT BIBLIOGRAPHICAL INDEX



```
                 COULOMB EXCHANGE ENERGY FROM SHELL-MODEL WAV      1719
                         EXCITATION OF PROTONS IN HELIUM II B      0011
OF ATOMIC AND MOLECULAR EXCITATION BY A TRAPPED-ELECTRON ME       0150
                 THERMAL EXCITATIONS IN LIQUID HE3.                1465
ENERGIES OF GROUND AND EXCITED NUCLEAR CONFIGURATIONS IN TH       0452
                         EXCITED STATES OF V51 AND CR53.           1691
                  4-PLUS EXCITED STATE IN OSMIUM-188.              1717
 NTERNAL PHOTOEFFECT AND EXCITON DIFFUSION IN CADMIUM AND ZIN      0123
OF THE CONTRIBUTION OF EXCITONS TO THE COMPLEX DIELECTRIC         1555
                 THERMAL EXPANSION OF SOME CRYSTALS WITH THE       0136
         ENERGY LEVELS IN F18 FROM THE N14/ALPHA,ALPHA/N14 AND     0547
 ON FROM AL27-PLUS-P AND F19-PLUS-P.                               0239
 TIC MEASUREMENTS OF THE FE-CR SPINELS.                            1603
                  BARIUM FERRATE III.                              0326
  MAGNETOSTATIC MODES IN FERRIMAGNETIC SPHERES.                    0059
             NICKEL-IRON FERRITE.                                  0397
       TRANSITION TO THE FERROELECTRIC STATE IN BARIUM TITANA      0413
   SUPERCONDUCTIVITY AND FERROMAGNETISM IN ISOMORPHOUS COMPOU      0089
 INTERPLANETARY MAGNETIC FIELD AND ITS CONTROL OF COSMIC-RAY       0589
                MAGNETIC FIELD DEPENDENCE OF ULTRASONIC ATTEN      0080
           RELATIVISTIC FIELD THEORY OF UNSTABLF PARTICLES.        0283
                 QUANTUM FIELD THEORIES WITH COMPOSITE PARTIC      0669
  A GENERALLY CONVARIANT FIELD THEORY.                             1826
 AND SURFACE STATES FROM FIELD-INDUCED CHANGES IN SURFACE REC      0369
 NGULAR DISTRIBUTIONS IN FISSION INDUCED BY ALPHA PARTICLES,       0536
 UTRON CROSS SECTIONS OF FISSIONABLE NUCLEI.                       0203
 AL COSMIC-RAY INTENSITY FLUCTUATIONS OBSERVED AT SOUTHERN ST      1798
                         FLUX OF COSMIC-RAY PARTICLES WITH Z-      0597
 NEUTRINO CORRELATION IN FORBIDDEN BETA DECAY.                     0244
                         FOURIER COEFFICIENTS OF CRYSTAL POTE      0073
 RVATION IN THE DECAY OF FREE AND BOUND LAMBDA PARTICLES.          0605
            STEADY-STATE FREE PRECESSION IN NUCLEAR MAGNETIC       1693
                         FREQUENCY SHIFT OF THE ZERO-FIELD HY      0449
                DECAY OF GADOLINIUM-159.                           0262
                         GAMMA RADIATION FROM AL27-PLUS-P AND      0239
 ECTIONAL CORRELATION OF GAMMA RAYS IN GE72.                       0229
 CISION DETERMINATION OF GAMMA RAYS FOLLOWING P,P-PRIME-GAMMA      0532
                         GAMMA-RAY THRESHOLD METHOD AND THE O      0461
 P/S32 AND S32/P,P-PRIME GAMMA/S32.                                1702
 ONSTANT OF YTTRIUM IRON GARNET AT 0 DEG K.                        0395
              LORENTZIAN GAS AND HOT ELECTRONS.                    1567
 TIBILITY OF AN ELECTRON GAS AT HIGH DENSITY.                      0328
 UCTIVITY OF AN ELECTRON GAS IN A GASEOUS PLASMA.                  0001
 OF AN ELECTRON GAS IN A GASEOUS PLASMA.                           0001
 DUCED BY VARIOUS BUFFER GASES.                                    0449
                  BUFFER GASES.                                    0450
                 IONIZED GAS.                                      1441
 EZORESISTANCE IN N-TYPE GA,AS.                                    1533
  IN ELECTRON-IRRADIATED GE AT 80 DEG K.                           0362
 LATION OF GAMMA RAYS IN GE72.                                     0229
 NERAL RELATIVITY AS THE GENERATORS OF COORDINATE TRANSFORMAT      0287
 ETORESISTANCE IN N-TYPE GERMANIUM AT LOW TEMPERATURES.            0317
 CONDUCTION ELECTRONS IN GERMANIUM.                                0298
 IATIVE RECOMBINATION IN GERMANIUM.                                0330
 PARTICLES IN LINEARIZED GRAVITATIONAL THEORY.                     0674
             ENERGIES OF GROUND AND EXCITED NUCLEAR CONFIGURA      0452
                         GROUND STATE OF TWO-ELECTRON ATOMS.       1649
           KINEMATICS OF GROWING WAVES.                            1488
 RIC CONSTANTS OF ALKALI HALIDE CRYSTALS.                          0090
                         HALL EFFECT, MAGNETORESISTANCE, AND       0381
                     TWO HALL EFFECTS OF IRON-COBALT ALLOYS.       1516
                         HALL MOBILITY OF CARRIERS IN IMPURE       0044
 A DILUTE BOSE SYSTEM OF HARD SPHERES. I. EQUILIBRIUM PROPERT      1419
 OLUME ANOMALY OF LIQUID HE3 ARISING FROM ITS NUCLEAR SPIN SY      1483
 L EXCITATIONS IN LIQUID HE3.                                      1465
  OF 95-MEV PROTONS WITH HE4.                                      1658
                SPECIFIC HEAT OF LI,F AND KI AT LOW TEMPERATU      0049
 TION OF DONOR STATES IN HEAT-TREATED SILICON.                     1546
 UCLEAR ENERGY LEVELS IN HEAVY ELEMENTS.                           0518
 XCITATION OF PROTONS IN HELIUM II BY COLD NEUTRONS.               0011
      MAGNETIC MOMENT OF HELIUM IN ITS 3S1 METASTABLE STATE.       1627
 LITY OF LI-PLUS IONS IN HELIUM.                                   0445
  OF SN, IN, TA, TL, AND HG.                                       0031
    ISOMERS IN TB158 AND HO163.                                    0489
     LORENTZIAN GAS AND HOT ELECTRONS.                             1567
 ICROWAVE PROPAGATION IN HOT MAGNETO-PLASMAS.                      1460
  OF THE ELECTRON ON THE HYDROGEN ENERGY LEVELS.                   1637
     DISSOCIATION OF THE HYDROGEN MOLECULE ION BY ELECTRON IM      0155
 SS OF SLOW ELECTRONS IN HYDROGEN.                                 0159
                         HYDROMAGNETIC EQUATIONS FOR TWO ISOT      1441
 SHIFT OF THE ZERO-FIELD HYPERFINE SPLITTING OF CS133 PRODUCE      0449
                  NARROW HYPERFINE ABSORPTION LINES OF CS133       0450
                         HYPERFINE-STRUCTURE SEPARATIONS AND       0186
 MASSES OF CHARGED SIGMA HYPERONS AND THE NEGATIVE K MESON.        0622
```

means. Such "Auto-Abstracts" may be derived from the text of a document by first compiling a word list, exclusive of "common words" such as articles, prepositions, conjunctions, etc. From this list a certain number of the highest ranking words would be assumed to be of high significance and be taken as a first criterion in the analysis of each of the sentences of the document. A second criterion would be how many of such words are present in a sentence and a third criterion would be how closely they are clustered among all the words in the sentence. By computing a "sentence significance factor" from these variables, a certain fraction of all sentences may be selected on the basis of these factors and be extracted from the document to form an Auto-Abstract.* A sample of such an Auto-Abstract is shown in Fig. 2.

Preparation of Word Patterns for Retrieval

One of the objectives of processing documents for retrieval is to reduce to a minimum the identifying elements needed to characterize documents adequately for a given application. It may therefore be expected that the means for accomplishing this differ widely with respect to the level of specificity desired. Information processing equipment is capable of preparing a variety of types of word patterns, suitable for various levels of retrieval requirements.

Keywords

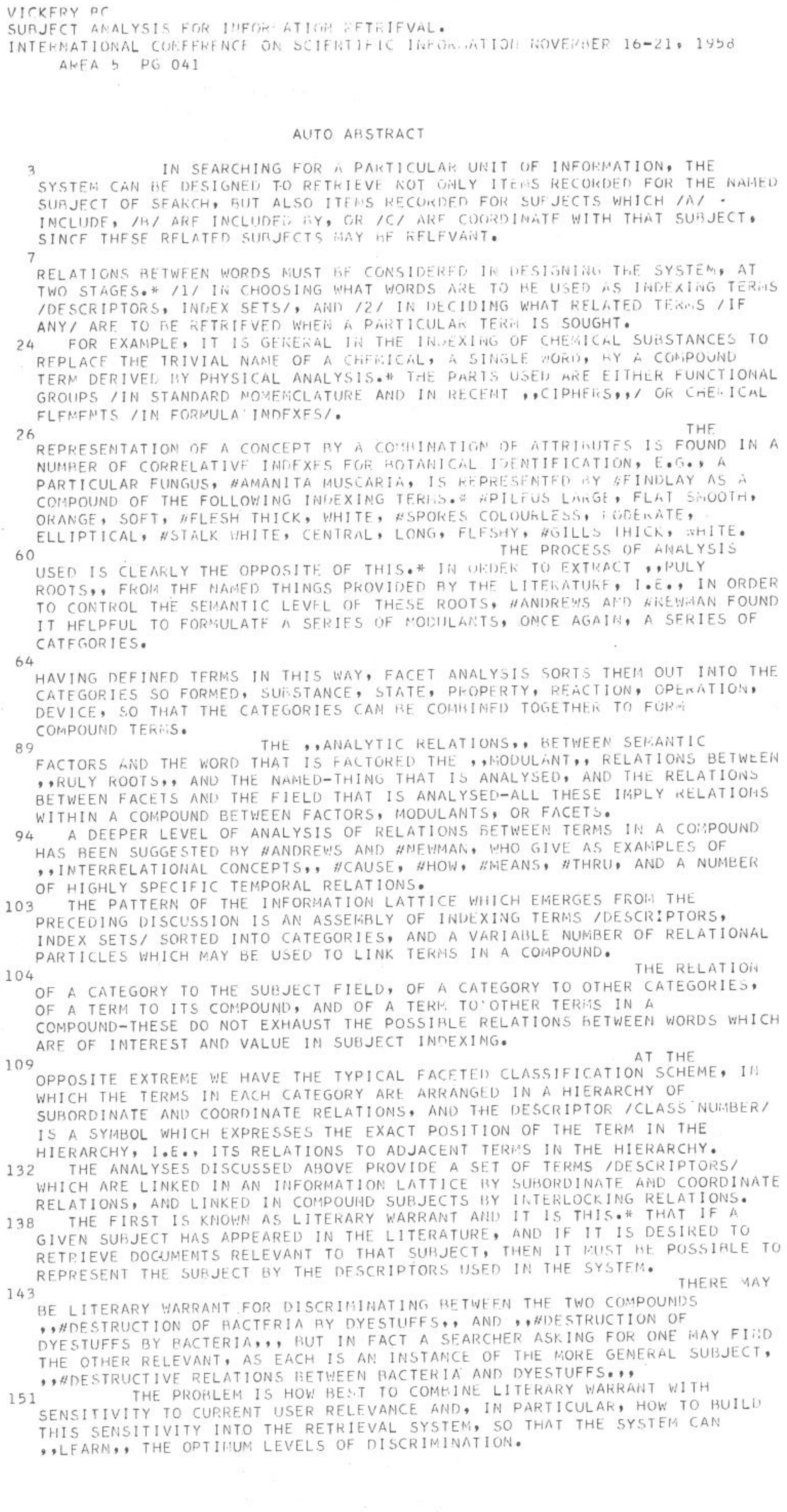
If frequency of occurrence is taken as a measure of word significance, a set of keywords may be derived from the word list compiled for a document as previously discussed. A limited portion of the highest ranking words of such a list may be selected to act as keywords. However, there is also a need for keywords whose significance is not necessarily dependent on frequency of usage. Such words may nevertheless be selected in addition by way of table look-up from a predetermined list of special words. In the case of proper names, these may be selected by recognizing the capitalized initial letter starting the words of this group.

Weighted Vocabularies

It may be argued that in a specific field of scientific endeavor a specific set of notions are used (Technese) and that the vocabulary of

* H. P. Luhn, "The Automatic Creation of Literature Abstracts, IBM Journal of Research and Development, April 1958. See also, "An Experiment in Auto-Abstracting", Progress Report, IBM Research Center, Yorktown Heights, N. Y. 1958, and T. R. Savage, "The Preparation of Automatic Abstracts on the IBM 704 Data Processing System", IBM Research Center, Yorktown Heights, N. Y., 1958

Fig. 2



```
VICKERY BC
SUBJECT ANALYSIS FOR INFORMATION RETRIEVAL.
INTERNATIONAL CONFERENCE ON SCIENTIFIC INFORMATION NOVEMBER 16-21, 1958
     AREA 5  PG 041


                        AUTO ABSTRACT

 3            IN SEARCHING FOR A PARTICULAR UNIT OF INFORMATION, THE
  SYSTEM CAN BE DESIGNED TO RETRIEVE NOT ONLY ITEMS RECORDED FOR THE NAMED
  SUBJECT OF SEARCH, BUT ALSO ITEMS RECORDED FOR SUBJECTS WHICH /A/ ·
  INCLUDE, /B/ ARE INCLUDED BY, OR /C/ ARE COORDINATE WITH THAT SUBJECT,
  SINCE THESE RELATED SUBJECTS MAY BE RELEVANT.
 7
  RELATIONS BETWEEN WORDS MUST BE CONSIDERED IN DESIGNING THE SYSTEM, AT
  TWO STAGES.* /1/ IN CHOOSING WHAT WORDS ARE TO BE USED AS INDEXING TERMS
  /DESCRIPTORS, INDEX SETS/, AND /2/ IN DECIDING WHAT RELATED TERMS /IF
  ANY/ ARE TO BE RETRIEVED WHEN A PARTICULAR TERM IS SOUGHT.
24   FOR EXAMPLE, IT IS GENERAL IN THE INDEXING OF CHEMICAL SUBSTANCES TO
  REPLACE THE TRIVIAL NAME OF A CHEMICAL, A SINGLE WORD, BY A COMPOUND
  TERM DERIVED BY PHYSICAL ANALYSIS.* THE PARTS USED ARE EITHER FUNCTIONAL
  GROUPS /IN STANDARD NOMENCLATURE AND IN RECENT ,,CIPHERS,,/ OR CHEMICAL
  ELEMENTS /IN FORMULA INDEXES/.
26                                                                 THE
  REPRESENTATION OF A CONCEPT BY A COMBINATION OF ATTRIBUTES IS FOUND IN A
  NUMBER OF CORRELATIVE INDEXES FOR BOTANICAL IDENTIFICATION, E.G., A
  PARTICULAR FUNGUS, #AMANITA MUSCARIA, IS REPRESENTED BY #FINDLAY AS A
  COMPOUND OF THE FOLLOWING INDEXING TERMS.* #PILFUS LARGE, FLAT SMOOTH,
  ORANGE, SOFT, #FLESH THICK, WHITE, #SPORES COLOURLESS, MODERATE,
  ELLIPTICAL, #STALK WHITE, CENTRAL, LONG, FLESHY, #GILLS THICK, WHITE.
60                                                THE PROCESS OF ANALYSIS
  USED IS CLEARLY THE OPPOSITE OF THIS.* IN ORDER TO EXTRACT ,,RULY
  ROOTS,, FROM THE NAMED THINGS PROVIDED BY THE LITERATURE, I.E., IN ORDER
  TO CONTROL THE SEMANTIC LEVEL OF THESE ROOTS, #ANDREWS AND #NEWMAN FOUND
  IT HELPFUL TO FORMULATE A SERIES OF MODULANTS, ONCE AGAIN, A SERIES OF
  CATEGORIES.
64
  HAVING DEFINED TERMS IN THIS WAY, FACET ANALYSIS SORTS THEM OUT INTO THE
  CATEGORIES SO FORMED, SUBSTANCE, STATE, PROPERTY, REACTION, OPERATION,
  DEVICE, SO THAT THE CATEGORIES CAN BE COMBINED TOGETHER TO FORM
  COMPOUND TERMS.
89                          THE ,,ANALYTIC RELATIONS,, BETWEEN SEMANTIC
  FACTORS AND THE WORD THAT IS FACTORED THE ,,MODULANT,, RELATIONS BETWEEN
  ,,RULY ROOTS,, AND THE NAMED-THING THAT IS ANALYSED, AND THE RELATIONS
  BETWEEN FACETS AND THE FIELD THAT IS ANALYSED-ALL THESE IMPLY RELATIONS
  WITHIN A COMPOUND BETWEEN FACTORS, MODULANTS, OR FACETS.
94   A DEEPER LEVEL OF ANALYSIS OF RELATIONS BETWEEN TERMS IN A COMPOUND
  HAS BEEN SUGGESTED BY #ANDREWS AND #NEWMAN, WHO GIVE AS EXAMPLES OF
  ,,INTERRELATIONAL CONCEPTS,, #CAUSE, #HOW, #MEANS, #THRU, AND A NUMBER
  OF HIGHLY SPECIFIC TEMPORAL RELATIONS.
103   THE PATTERN OF THE INFORMATION LATTICE WHICH EMERGES FROM THE
  PRECEDING DISCUSSION IS AN ASSEMBLY OF INDEXING TERMS /DESCRIPTORS,
  INDEX SETS/ SORTED INTO CATEGORIES, AND A VARIABLE NUMBER OF RELATIONAL
  PARTICLES WHICH MAY BE USED TO LINK TERMS IN A COMPOUND.
104                                                           THE RELATION
  OF A CATEGORY TO THE SUBJECT FIELD, OF A CATEGORY TO OTHER CATEGORIES,
  OF A TERM TO ITS COMPOUND, AND OF A TERM TO OTHER TERMS IN A
  COMPOUND-THESE DO NOT EXHAUST THE POSSIBLE RELATIONS BETWEEN WORDS WHICH
  ARE OF INTEREST AND VALUE IN SUBJECT INDEXING.
109                                                              AT THE
  OPPOSITE EXTREME WE HAVE THE TYPICAL FACETED CLASSIFICATION SCHEME, IN
  WHICH THE TERMS IN EACH CATEGORY ARE ARRANGED IN A HIERARCHY OF
  SUBORDINATE AND COORDINATE RELATIONS, AND THE DESCRIPTOR /CLASS NUMBER/
  IS A SYMBOL WHICH EXPRESSES THE EXACT POSITION OF THE TERM IN THE
  HIERARCHY, I.E., ITS RELATIONS TO ADJACENT TERMS IN THE HIERARCHY.
132   THE ANALYSES DISCUSSED ABOVE PROVIDE A SET OF TERMS /DESCRIPTORS/
  WHICH ARE LINKED IN AN INFORMATION LATTICE BY SUBORDINATE AND COORDINATE
  RELATIONS, AND LINKED IN COMPOUND SUBJECTS BY INTERLOCKING RELATIONS.
138   THE FIRST IS KNOWN AS LITERARY WARRANT AND IT IS THIS.* THAT IF A
  GIVEN SUBJECT HAS APPEARED IN THE LITERATURE, AND IF IT IS DESIRED TO
  RETRIEVE DOCUMENTS RELEVANT TO THAT SUBJECT, THEN IT MUST BE POSSIBLE TO
  REPRESENT THE SUBJECT BY THE DESCRIPTORS USED IN THE SYSTEM.
143                                                             THERE MAY
  BE LITERARY WARRANT FOR DISCRIMINATING BETWEEN THE TWO COMPOUNDS
  ,,#DESTRUCTION OF BACTERIA BY DYESTUFFS,, AND ,,#DESTRUCTION OF
  DYESTUFFS BY BACTERIA,,, BUT IN FACT A SEARCHER ASKING FOR ONE MAY FIND
  THE OTHER RELEVANT, AS EACH IS AN INSTANCE OF THE MORE GENERAL SUBJECT,
  ,,#DESTRUCTIVE RELATIONS BETWEEN BACTERIA AND DYESTUFFS.,,
151          THE PROBLEM IS HOW BEST TO COMBINE LITERARY WARRANT WITH
  SENSITIVITY TO CURRENT USER RELEVANCE AND, IN PARTICULAR, HOW TO BUILD
  THIS SENSITIVITY INTO THE RETRIEVAL SYSTEM, SO THAT THE SYSTEM CAN
  ,,LEARN,, THE OPTIMUM LEVELS OF DISCRIMINATION.
```

the language expressing these notions is comparatively small and distinct with respect to other such languages. A word list could therefore be compiled from a set of documents most representative of the specific field and each word be given a weight depending on frequency of occurrence. The characterization of a document for retrieval may then be accomplished by recording a rather substantial portion of the vocabulary for each document. Retrieval could be based on the degree of correlation between such vocabularies.

Word Pairs

More specific terms for characterizing a document may be derived from word pairs. The assumption here is that the probability is high that words appearing close to each other in a sentence modify or supplement each other or are specifically related in various other ways. Such pairs may be automatically extracted from text on the basis of frequency of occurrence and degree of proximity, a measure previously mentioned for preparing auto-abstracts.

A word pattern for retrieval may consist of word pairs selected on the basis of frequency of occurrence or other measurements that may be performed by the machine

Word pairs may be compiled and tabulated by machine in the form of a word pair matrix as illustrated in Fig. 3. This format is useful for analytical work. The recording of the word pair pattern for retrieval may consist of a list of the pairs in a given order. Another form of listing may be obtained by the node and branch method in which given words are followed each by a list of words they are paired with. *

Phrases

A more specific identification of word relationships may be desirable in certain applications. In this case it may become necessary to establish word associations more specifically in terms of syntactical units and to recognize parts of speech and their interaction. An account may then be given as to which word or words modify a given noun, for example. The analytical process to be performed by machine for this degree of identification are considerable and approach the techniques which are essential to machine translation. The question is what simplest process will give acceptable results. One method which is liable to produce useful units of meaning consists of the recognition and extraction

* For further details see: H. P. Luhn, "Auto-Encoding of Documents for Information Retrieval Systems", IBM Monograph, 1958.

Fig. 3

```
DOCUMENT NO IC12 CONTAINING 4068 WORDS

VICKERY BC
SUBJECT ANALYSIS FOR INFORMATION RETRIEVAL.
PREPRINTS OF PAPERS FOR ICSI 1958   IC12

          MATRIX OF WORD PAIRS
      GIVING FREQUENCY OF OCCURRENCE

     NUMBER OF
WRD  DIFFERENT
NUM  PAIRINGS
BER  PER WORD        WORD            WORD FREQUENCY

 01.   24     TERMS                       76
 02.   20    17 RELATION                    59
 03.   17     1 8 ANALYSIS                    54
 04.   12     . 1 . SYSTEM                      40
 05.   12     2 . .16 RETRIEVAL                   31
 06.   23     1 3 2 1 . SUBJECT                     30
 07.   13    14 . 1 2 . 4 INDEX                       27
 08.    8     . . 7 . 1 1 2 LEVEL                       26
 09.   13     9 2 . . . 2 2 . CATEGORISATION              23
 10.    9    12 . . . . 1 . . . COMPOUND                    20
 11.    8     . . . . 1 3 . . . . RELEVANCE                   20
 12.   11     2 1 1 2 . 3 . 4 . . 3 DISCRIMINATE                19
 13.   11     3 1 . 1 . 1 3 . 1 . . . DESCRIBED                   17
 14.    7     . 1 5 . . . . 8 . . . . . SEMANTIC                    17
 15.    8     1 . . . . . . . 1 . . . . . CATEGORIES                  16
 16.    3     2 . 6 . . 1 . . . . . . . . . DEFINITION                  15
 17.    6     . . . . 1 . 1 . 1 2 . . 2 . 2 FORM                        15
 18.    3     . . . 2 6 . . . . . . . . . 1 INFORMATION                 15
 19.    5     . 3 1 . . . 1 . . 1 . . 1 . . WORD                        15
 20.    3     3 5 . . . 3 . . . . . . . . . COORDINATE                  14
 21.    2     1 . . . . . 1 . . . . . . . . CLASSIFICATION              13
 22.    3     1 3 . . . . . . . . . . 1 . . HIERARCHICAL                13
 23.    3     1 . . . . 2 1 . . . . . . . . GENERAL                     12
 24.    2     1 1 . . . . . . . . . . . . . INCLUSION                   12
 25.    3     1 1 . 1 . . . . . . . . . . . LATTICE                     12
 26.    4     . . 2 2 . . . 1 1 . . . . . . OPERATE                     12
 27.    1     . . 3 . . . . . . . . . . . . PROCEDURE                   12
 28.    3     . . . . . . . . . 1 . 1 . 2 . FACT                        11
 29.    1     . . . . . 1 . . . . . . . . . LITERARY                    11
 30.    4     . 2 . . 1 . . . 1 . 1 . . . . PART                        11
 31.    3     . . . . . 1 . . . . . 1 . 2 . PROBLEM                     11
 32.    2     . . . . 1 2 . . . . . . . . . SEARCH                      11
 33.    4     . 4 . . . 1 . . 1 . . 1 . . . SPECFIC                     11
 34.    4     1 . 4 . . 1 . . . . . . . 1 . TECHNIQUE                   11
 35.    3     . . 1 . . 2 . . 1 . . . . . . FIELD                       10
 36.    2     3 . . . . . . . 2 . . . . . . ARRANGE                     09
 37.    2     . . 1 . . . . . . 1 . . . . . ATTRIBUTES                  09
 38.    1     . . . . . . 1 . . . . . . . . EQUALS                      09
 39.    3     1 1 3 . . . . . . . . . . . . FACE                        09
 40.    1     . 1 . . . . . . . . . . . . . NUMBER                      09
 41.    5     . 2 1 1 . 1 . . . . 1 . . . . POSSIBILITY                 09
 42.    3     1 . . . . . . 1 . 1 . . . . . PROVIDE                     09
 43.    1     . . . . . . . . . . 4 . . . . USER                        09
 44.    2     . 2 . . 1 . . . . . . . . . . BACTERIA                    08
 45.    3     . . 1 . . . . . . . . . 1 . 1 CONCEPT                     08
 46.    5     3 2 . . . . . . . 1 . . 1 1 . LINK                        08
 47.    2     . . . 1 1 . . . . . . . . . . MACHINE                     08
 48.    2     1 . . . . . . . . . 1 . . . . SCHEME                      08
 49.    6     1 . . 1 1 1 1 . . . . . . . 1 USE                         08
 50.    2     . . . . . 1 . . . . . 1 . . . WARRANT                     08
```

of "prepositional phrases" contained in the text.* Such phrases may be identified by the machine through look-up in a stored word list of prepositions. An arbitrarily fixed number of words following the preposition is then considered to constitute the wanted phrase. The non-common words contained in such phrases may be taken as truly modified expressions and be used to form retrieval patterns composed of pairs or groups of the words associated in this fashion.

If prepositional phrases are found to be specifically representative of the information content of a document it might be advantageous to derive basic word lists from these phrases only instead of the complete text.

Compilation of Dictionaries and Thesauri

In the previous discussions no reference has been made to the fact that variation of word usage might interfere with the utility of patterns composed of the words as found in a document. It is important, therefore, that due consideration be given to this situation and means be provided to overcome such variations by some process of normalization. Just as an author or reader may turn to a dictionary to clarify the meaning and usage of a given word, so it will be necessary for a machine to resolve variation of word usage with the aid of a device the functions of which resemble that of a dictionary at one level and of a thesaurus at another level of requirement.

The compilation of special dictionaries and thesauri is an intellectual task calling for decisions on the basis of complete familiarity with the given field. Such work may however be significantly simplified and superior results be obtained if full use is made of information processing equipment for organizing and presenting the material in a manner which will bring out the points on which decisions will have to depend.

The statistical material that may be required in the manual compilation of dictionaries and thesauri may be derived from the original texts in any desired form and degree of detail. This is also true for supplementary material needed for periodic adjustments and updating. This latter material may be supplied currently as a by-product of the encoding procedure for each new document.

Additional statistical material of interest may be derived from the retrieval functions of a system and may serve to evaluate the effectiveness of the encoding structures employed.

* P. B. Baxendale, "Machine-Made Index for Technical Literature - An Experiment", IBM Journal of Research and Development, October 1958.

Dictionaries and thesauri are made accessable to the encoding process by storing them in the machine. Words of the text would be looked up as a matter of course to obtain their normalized version either in the form of another word or in the form of a code word or number.

Derivation of Code Words

In many systems it is desirable to reduce lengthy expressions into more compact codes, thereby saving storage space and processing time. Such codes may be derived from the original notations by systematic reduction procedures readily performed by machines and may be stored in the form of code dictionaries or be applied as part of the encoding procedure. Typical examples of such codes are given in Fig. 4.

Statistical Analysis of Texts

The various schemes enumerated in this paper are based on the capabilities of machines to analyze textual material in many ways. Once certain basic operations have been performed on the text, such as sorting of all words in alphabetic order, it takes comparatively little effort to derive additional statistical information useful not only for the encoding process proper but also for the overall design of a system, its supervision and upkeep. Much of this information is of a kind which would be entirely impractical or well-nigh impossible to obtain with manual encoding operations.

By way of illustration, there is shown in Fig. 5 on the following pages the machine print-out of statistical information derived from the text of a typical scientific document. This information consists of 14 lists and tables relating to various properties and relationships of the words in the document. The headings preceding each list or table are self-explanatory.

Fig. 4

DERIVATION BY MACHINE OF 4-LETTER CODE WORDS BY THE SIGNIFICANT LETTER SPELLING METHOD. *

ABSTRACT	ABRC	ANALOGIES	ANLG	BACTERIA	BACR
ABSTRACTING	ABRC	ANALOGY	ANLG	BASAL	BAS
ABSTRACTOR	ABRC	ANIMALS	ANIM	BASED	BAS
ABSTRACTS	ABRC	ANSWER	ANSW	BASIC	BAS
ACADEMIC	ACDM	APPROACH	APCH	BATTEN	BATT
ACCEPT	ACPT	APPROPRIATE	APPR	BELIEF	BLIF
ACCOMPLISH	ACMP	APPROVAL	APRV	BINDING	BIND
ACCORDANCE	ACRD	APPROXIMATE	APXM	BIRTH	BIRH
ACCORDING	ACRD	AREA	AREA	BOARD	BORD
ACTUAL	ACTU	ARRANGE	ARNG	BOND	BOND
ADDRESS	ADRS	ARRANGED	ARNG	BOOK	BOOK
ADJECTIVAL	AJCV	ARRANGING	ARNG	BROAD	BROD
ADJUNCTS	AJUC	ASCERTAIN	ASCR	BUDGETED	BUDG
ADOPT	ADOP	ASKED	ASK	BUILDER	BULD
ADULT	ADUL	ASPECT	ASPC	BULLETIN	BULT
ADVANCE	ADVN	ASSEMBLE	ASMB	CALCULATE	CCUL
ADVANTAGE	ADVG	ASSIGNED	ASIG	CAMEL	CAML
AERODYNAMIC	ADYM	ASSOCIATE	ASCI	CANONICAL	CNON
AGE	AGE	ASSOCIATED	ASCI	CAP	CAP
AIR	AIR	ASSOCIATION	ASCI	CARBON	CRBN
ALGEBRA	ALGB	ASSUME	ASUM	CARD	CARD
ALGORITHM	ALGM	ATOM	ATOM	CASE	CASE
ALPHABET	ALPB	ATTENDED	ATND	CELL	CELL
ALTERNATE	ALRN	ATTRIBUTES	ARBU	CENTER	CNTR
ALUMINUM	AMUM	AUTHOR	AUTH	CENTERS	CNTR
AMBIGUITIES	AMBG	AUTHORS	AUTH	CENTRAL	CNTR
AMBIGUITY	AMBG	AUTOMATIC	AUOM	CHAMBER	CHMB
AMERICA	AMRC	AUXILIARY	AUXL	CHANCE	CHAN
AMERICAN	AMRC	AVERAGE	AVRG	CHANGE	CHNG
AMOUNT	AMOU	BACKGROUND	BKGU	CHARGE	CHRG

* For details see: H. P. Luhn, "Superimposed Coding with the Aid of Randomizing Squares for Use in Mechanical Information Searching Systems", Chapter 23 in "Punched Cards", 2nd Edition, Reinhold Publishing Corp., New York, 1958

DERIVATION BY MACHINE OF 11-CHARACTER INDEX CODES FOR THE IDENTIFICATION OF BIBLIOGRAPHICAL ITEMS.

```
CCGOML-52-WHT  C.C.GOODRICH MEMORIAL LIBRARY
                 WHY AND HOW THE TECHNICAL LIBRARY SHOULD BE SET UP AND
                   UTILIZED IN CREATIVE ENGINEERING.
                 MACHINE DESIGN SEPT 1952 PP. 111
HOLMJE-57-MDD  HOLMSTROM JE
                 MULTILINGUAL DICTIONARIES AND DOCUMENTATION
                 NACHRICHTEN DOKUMENTATION MAR. 1957
INSTAS----SST  INSTITUTE OF THE AERONAUTICAL SCIENCES
                 SYMPOSIUM ON STANDARDIZATION IN TECHNICAL INFORMATION
                   SERVICES FOR GOVERNMENT
                 US RESEARCH AND DEVELOPMENT BOARD
JOHNHU-55-MIP  JOHNS HOPKINS UNIVERSITY
                 MEDICAL INDEXING PROJECT, FINAL REPORT.
                 WELCH MEDICAL LIBRARY,JOHNS HOPKINS UNIVERSITY% MEDICAL
                   INDEXING PROJECT, FINAL REPORT, 1955
KENTA -57-MSM  KENT A
                 MACHINE SEARCHING OF METALLURGICAL LITERATURE.
                 METAL PROGRESS, FEB. 1957
KINGGW-55-NAI  KING GW
                 A NEW APPROACH TO INFORMATION STORAGE.
                 CONTROL ENGINEERING AUGUST 1955
KOELGJ-58-PFM  KOELEWIJN GJ
                 THE POSSIBILITIES OF FAR-REACHING MECHANIZATION OF NOVELTY
                   SEARCH OF THE PATENT LITERATURE.
                 PREPRINTS OF PAPERS FOR THE INTERNATIONAL CONFERENCE ON
                   SCIENTIFIC INFORMATION WASH. DC 1958
MAC CG-54-CFS  MAC CASLAND GE
                 A CONCISE FORM FOR SCIENTIFIC LITERATURE CITATIONS.
                 SCIENCE 120, JULY 1954
MIDWRI-57-EBM  MIDWEST RESEARCH INSTITUTE, KANSAS CITY,MO.
                 ELECTRONIC BRAIN MULLS NEW CHEMICAL USES.
                 CHEMICAL WEEK NOV. 23, 1957
NATLBS-57-SPE  NATL.BUR. OF STANDARDS WASHINGTON DC
                 SYNTAX PATTERNS IN ENGLISH STUDIED BY ELECTRONIC COMPUTER.
                 COMPUTERS AND AUTOMATION  JULY 1957
```

Note: The letters or numbers extracted by the machine to form the code have been underlined.

Fig. 5

STATISTICAL ANALYSIS

```
AXELROD J, NATL. INST. OF MENTAL HEALTH, BETHESDA MARYLAND
PRESENCE, FORMATION, AND METABOLISM OF NORMETANEPHRINE IN THE BRAIN
R283  SCIENCE APRIL 4, 1958 VOLUME 127 NUMBER 3301 PS 754 PE 755

LIST OF NON-COMMON WORDS IN ALPHABETIC ORDER WITH INDICATION
OF CONSOLIDATED WORDS AND WORD LOCATION IN TEXT

FREQ  DOC       NO  WORD                  LOCATIONS IN TEXT

  1   R283       1 0.30                   633
  2   R283       2 1957                   747    763
  2   R283       3 10.                    200    304
  1   R283       4 100                    795
  1   R283       5 2008                   271
  1   R283       6 2000                   284
  4   R283       7 3-METHOXY-4 HYDROX     541    641    664    688
  1   R283       8 6-DICHLOROQUINONE      296
  1   R283       9 ABILITY                503
  1   R283      10 ABSENCE                426
  3   R283      11 ACETATE                590    596    281
      R283      11       ACETIC
  7   R283      12 ACID                   233    282    542    642    665    689    561
      R283      12       ACIDIFIED
  2   R283      13 ACTION                  37     58
  1   R283      14 ADJUSTED               197
  1   R283      15 ADULT                  139
  1   R283      16 AGENT                    3
  2   R283      17 ALCOHOL                213    299
  1   R283      18 ALDEHYDE               549
  2   R283      19 AMINE                   33    486
  1   R283      20 AMMONIA                268
  1   R283      21 AMOUNTS                483
  1   R283      22 AQUEOUS                584
  1   R283      23 AREAS.1                806
  1   R283      24 ARMY                   802
  1   R283      25 ASCENDING              261
  2   R283      26 AUTHENTIC              319    638
  1   R283      27 BICARBONATE            580
  1   R283      28 BLOCK                  161
  3   R283      29 BLUE                   311    340    625
  2   R283      30 BORATE                 202    305
 16   R283      31 BRAIN                   26    176    355    374    393    413    466
      R283      31                        480    505    530    547    646    679    734
      R283      31                        180    365
      R283      31       BRAINS
  1   R283      32 BROKE                  783
  2   R283      33 BUFFER                 203    306
  2   R283      34 BULLETIN               738    754
  1   R283      35 BUTANOL                572
  1   R283      36 CATECHOL                32
  3   R283      37 CENTRAL                 14     40    705
  1   R283      38 CENTRIFUGATION         221
  1   R283      39 CHILLED                183
  3   R283      40 CHLORIDE               415    297    621
      R283      40       CHLORIMIDE
  5   R283      41 CHROMATOGRAM           292    616    446    260    610
      R283      41       CHROMATOGRAPHED   CHROMATOGRAPHY
  1   R283      42 CLARIFICATION          215
  2   R283      43 COLOR                  333    630
  1   R283      44 COMMUNIST              775
  2   R283      45 COMPOUND               384    658
  3   R283      46 CONCERNED              394     47    694
      R283      46       CONCERNING
  1   R283      47 CONCLUSIVE             692
  1   R283      48 CONJECTURAL             45
  1   R283      49 CONSIDERABLE           482
  1   R283      50 CONSTITUTES            722
  1   R283      51 DAILY                  158
  1   R283      52 DAYS                   169
  5   R283      53 DEAMINATED             528    537     69    389    685
      R283      53       DEAMINATION
  1   R283      54 DECAPITATED            177
  1   R283      55 DEHYDROGENASE          550
  5   R283      56 DESCRIBED              448    612    668    713    118
      R283      56       DESCRIBES
  1   R283      57 DETECTED               363
  1   R283      58 DICHLOROQUINONE        620
  1   R283      59 DIPHOSPHOPYRIDINE      552
  1   R283      60 DISAPPEARANCE          490
  1   R283      61 DISTINCT               310
  1   R283      62 D-BITARTRATE           406
  1   R283      63 DONOR                  115
  1   R283      64 DRYNESS                238
  1   R283      65 ELSEWHERE              716
  1   R283        66 ENDOGENOUS           376
  2   R283        67 ENZYMATICALLY        441    539
  2   R283        68 ENZYMES              131    390
  1   R283        69 ESTIMATED            344
  2   R283        70 ETHYL                599    595
  2   R283        71 EVAPORATED           236    246
  2   R283        72 EVIDENCE             524    693
  1   R283        73 EXPERIMENTS          712
  6   R283        74 EXTRACT              219    234    255    597    205    565
      R283        74        EXTRACTED
  1   R283        75 FAINT                308
  1   R283        76 FILTER               265
  2   R283        77 FOLLOWED             300    672
      R283        77        FOLLOWING
  3   R283        78 FORMATION            135    421    655
  1   R283        79 FORMED               442
  1   R283        80 FOUR                 138
  4   R283        81 FRACTION             410    437    501    650
  1   R283        82 GRAM                 353
  1   R283        83 HAVING               659
  3   R283        84 HC1                  193    231    563
  1   R283        85 HIGHLY                10
  1   R283        86 HOMOGENATE           195
  1   R283        87 HOMOGENIZED          187
  1   R283        88 HOURS                556
  1   R283        89 IMMEDIATELY          186
  1   R283        90 IMPORTANT            724
  5   R283        91 INCUBATED            478    497    643    403    544
      R283        91        INCUBATING       INCUBATION
  1   R283        92 INDICATED            670
  1   R283        93 INDUSTRIAL           792
  2   R283        94 INHIBITOR            154    515
  2   R283        95 INSTITUTE            739    755
  1   R283        96 INTENSITY            337
  1   R283        97 INTRAPERITONEALLY    159
  2   R283        98 INVOLVED             132    100
      R283        98        INVOLVES
  3   R283        99 IPRONIAZID           149    173    516
  1   R283       100 IPRONIAZID-TREATED   473
  2   R283       101 ISOAMYL              212    218
  1   R283       102 ISOLATED             444
  1   R283       103 ISOPROPANOL          267
  2   R283       104 JULY                 746    762
  1   R283       105 KILOGRAM             156
  2   R283       106 KNOWLEDGE             46     20
      R283       106        KNOWN
  1   R283       107 LABORATORY            85
  1   R283       108 LACKING              709
  1   R283       109 LATTER               383
  2   R283       110 LAYER                573    585
  1   R283       111 LIGHT                 53
  1   R283       112 L-MOREPINEPHRINE     405
  1   R283       113 LOCALIZED             11
  1   R283       114 MAGNESIUM            414
  1   R283       115 MALE                 140
  1   R283       116 MANNER               327
  1   R283       117 MARKEDLY             521
 12   R283       118 METABOLIC            398    673     23     49     94    137    164
      R283       118                      700    728    518    386    488
      R283       118        METABOLISM      METABOLIZE        METABOLIZED
  3   R283       119 METHANOL             244    254    114
      R283       119        METHYL
  2   R283       120 MILLION              791    796
  6   R283       121 MITOCHONDRIA         431    481    506    531    548    647
  1   R283       122 MIXTURE              559
  1   R283       123 M-O-METHYLNOREPINE   105
  1   R283       124 MODE                  56
  2   R283       125 MONOAMINE            152    513
  3   R283       126 NERVOUS               15     41    706
  1   R283       127 NEUROHUMORAL           2
  2   R283       128 NITROGEN             252    605
  2   R283       129 N-BUTANOL            280    570
  2   R283       130 NO.7                 749    765
  9   R283       131 NOREPINEPHRINE         4     62     96    377    463    676    680
      R283       131                      702    731
  1   R283       132 NORMAL               471
 16   R283       133 NORMETANEPHRINE      104    122    166    322    351    360    379
      R283       133                      423    443    476    492    519    526    546
      R283       133                      644    684
  1   R283       134 NUCLEOTIDE           553
  2   R283       135 OBSERVATIONS         370    667
  1   R283       136 OBTAINED             533
  2   R283       137 OCCURRED             440    493
  2   R283       138 O.05                 227    230
      R283       138        O.05N
  2   R283       139 O.1                  294    346
  1   R283       140 O.1M                 302
  1   R283       141 O.1N                 192
```

```
1   R283    142 0.2                       348
1   R283    143 0.45                      316
1   R283    144 0.50                      314
1   R283    145 0.60                      635
5   R283    146 O-METHYLATION             101    439    461    681    721
1   R283    147 0086                      735
1   R283    148 OSBORNE-MENDEL            142
3   R283    149 OXIDASE                   153    514    687
    R283    149       OXIDATION
1   R283    150 OXIDIZING                 535
1   R283    151 PAPER                     266
1   R283    152 PARTY                     776
1   R283    153 PAST                       61
4   R283    154 PATHWAY                    91    674    697     77
    R283    154       PATHWAYS
2   R283    155 PEASANTRY                 779    797
    R283    155       PEASANTS
3   R283    156 PERCENT                   270    295    578
1   R283    157 PHOSPHATE                 150
1   R283    158 PLAYS                      34
1   R283    159 POOLED                    184
1   R283    160 POSSIBILITY                72
1   R283    161 PREPARED                  507
1   R283    162 PRETREATED                510
2   R283    163 PRINCIPAL                  90    696
1   R283    164 PROCESSES                 399
1   R283    165 PRODUCT                   538
1   R283    166 R364                      737
4   R283    167 RAT                        99    125    373    412
1   R283    168 RATE                      459
5   R283    169 RATS                      141    175    368    474    509
3   R283    170 REACTION                  109    334    558
1   R283    171 REAGENT                   622
1   R283.   172 RECENT                     81
1   R283    173 RECOGNIZED                 80
2   R283    174 REDUCED                   522    599
3   R283    175 REEXTRACTED               224    575    587
1   R283    176 REMOVED                   182
1   R283    177 REPORT                    117
1   R283    178 REQUIRES                  110
3   R283    179 RESULTED                  418    652    674
    R283    179       RESULTING
1   R283    180 REVOLUTION                782
5   R283    181 RF-S                      313    331    453    632    662
1   R283    182 ROLE                       29
1   R283    183 ROUTE                     725
1   R283    184 RURAL                     805
1   R283    185 RUSSIA                    786
3   R283    186 SAMPLE                    320    457    639
1   R283    187 SHED                       51
4   R283    188 S-ADENOSYLMETHIONI        111    417    428    683
1   R283    189 SOCIETY                   771
1   R283    190 SODIUM                    579
4   R283    191 SOLUBLE                   409    436    500    649
1   R283    192 SOLUTION                  581
1   R283    193 SOLVENT                   277
1   R283    194 SOVIET                    770
3   R283    195 SPOT                      312    341    626
2   R283    196 SPRAYED                   618    289
    R283    196       SPRAYING
1   R283    197 STOCK                     143
3   R283    198 STUDIED                   402    742    758
    R283    198       STUDY
2   R283    199 SUBJECTED                 257    607
1   R283    200 SUBSTITUTED               433
1   R283    201 SUGGEST                   371
1   R283    202 SYNTHETIC                 456
4   R283    203 SYSTEM                     16     42    278    707
1   R283    204 TARR                      772
1   R283    205 TECHNIQUE                 262
1   R283    206 TISSUE                    356
2   R283    207 TRANSFORMED                67    375
    R283    207       TRANSFORMS
2   R283    208 TREATED                   323    171
    R283    208       TREATMENT
2   R283    209 TWICE                     157    225
2   R283    210 TWO-DIMENSIONAL           259    609
1   R283    211 UNTREATED                 367
2   R283    212 USSR                      745    761
1   R283    213 VACUUM                    240
2   R283    214 VOL.IV                    748    764
6   R283    215 VOLUME                    190    228    250    603    210    568
    R283    215       VOLUMES
1   R283    216 WATER                     283
1   R283    217 WHATMAN-NO.               264
2   R283    218 WORK                       82    793
    R283    218       WORKERS
1   R283    219 YIELD                     103
```

```
DOC  R283        35 SENTENCES       30 CONSOLDTNS


LIST OF NON-COMMON WORDS IN DOCUMENT IN FREQUENCY ORDER WITH
ABSOLUTE AND RELATIVE FREQUENCY INDICATED
1DOC  R283

  16 .01985 BRAIN             NORMETANEPHRINE
  12 .01488 METABOLIC
   9 .01116 NOREPINEPHRINE
   7 .00868 ACID
   6 .00744 EXTRACT           MITOCHONDRIA      VOLUME
   5 .00620 CHROMATOGRAM      DEAMINATED        DESCRIBED
            INCUBATED         O-METHYLATION     RATS
            RF-S
   4 .00496 3-METHOXY-4-HYDROXFRACTION          PATHWAY
            RAT               S-ADENOSYLMETHIONISOLUBLE
            SYSTEM
   3 .00372 ACETATE           BLUE              CENTRAL
            CHLORIDE          CONCERNED         FORMATION
            HC1               IPRONIAZID        METHANOL
            NERVOUS           OXIDASE           PERCENT
            REACTION          REEXTRACTED       RESULTED
            SAMPLE            SPOT              STUDIED
   2 .00248 1957              10.               ACTION
            ALCOHOL           AMINE             AUTHENTIC
            BORATE            BUFFER            BULLETIN
            COLOR             COMPOUND          ENZYMATICALLY
            ENZYMES           ETHYL             EVAPORATED
            EVIDENCE          FOLLOWED          INHIBITOR
            .INSTITUTE        INVOLVED          ISOAMYL
            JULY              KNOWLEDGE         LAYER
            MILLION           MONOAMINE         NITROGEN
            N-BUTANOL         NO.7              OBSERVATIONS
            OCCURRED          0.05              0.1
            PEASANTRY         PRINCIPAL         REDUCED
            SPRAYED           SUBJECTED         TRANSFORMED
            TREATED           TWICE             TWO-DIMENSIONAL
            USSR              VOL.IV            WORK
   1 .00124 0.30              100               2008
            2000              6-DICHLOROQUINONE ABILITY
            ABSENCE           ADJUSTED          ADULT
            AGENT             ALDEHYDE          AMMONIA
            AMOUNTS           AQUEOUS           AREAS.1
            ARMY              ASCENDING         BICARBONATE
            BLOCK             BROKE             BUTANOL
            CATECHOL          CENTRIFUGATION    CHILLED
            CLARIFICATION     COMMUNIST         CONCLUSIVE
            CONJECTURAL       CONSIDERABLE      CONSTITUTES
            DAILY             DAYS              DECAPITATED
            DEHYDROGENASE     DETECTED          DICHLOROQUINONE
            DIPHOSPHOPYRIDINE DISAPPEARANCE     DISTINCT
            D-BITARTRATE      DONOR             DRYNESS
            ELSEWHERE         ENDOGENOUS        ESTIMATED
            EXPERIMENTS       FAINT             FILTER
            FORMED            FOUR              GRAM
            HAVING            HIGHLY            HOMOGENATE
            HOMOGENIZED       HOURS             IMMEDIATELY
            IMPORTANT         INDICATED         INDUSTRIAL
            INTENSITY         INTRAPERITONEALLY IPRONIAZID-TREATED
            ISOLATED          ISOPROPANOL       KILOGRAM
            LABORATORY        LACKING           LATTER
            LIGHT             L-NOREPINEPHRINE  LOCALIZED
            MAGNESIUM         MALE              MANNER
            MARKEDLY          MIXTURE           M-O-METHYLNOREPINE
            MODE              NEUROHUMORAL      NORMAL
            NUCLEOTIDE        OBTAINED          0.1M
            0.1N              0.2               0.45
            0.50              0.60              0086
            OSBORNE-MENDEL    OXIDIZING         PAPER
            PARTY             PAST              PHOSPHATE
            PLAYS             POOLED            POSSIBILITY
            PREPARED          PRETREATED        PROCESSES
            PRODUCT           R364              RATE
            REAGENT           RECENT            RECOGNIZED
            REMOVED           REPORT            REQUIRES
            REVOLUTION        ROLE              ROUTE
            RURAL             RUSSIA            SHED
            SOCIETY           SODIUM            SOLUTION
            SOLVENT           SOVIET            STOCK
            SUBSTITUTED       SUGGEST           SYNTHETIC
            TARR              TECHNIQUE         TISSUE
            UNTREATED         VACUUM            WATER
            WHATMAN-NO.       YIELD
```

COMMON WORD LIST

THE000	71	OF000	36	TO0000	15	A00000	6
IN0000	23	IS000	4	AND000	23	WE0000	
THAT00	8	WHICH		IT0000	4	FROM00	4
BY0000	6	ARE00		BE0000	3	AS0000	10
AN0000	3	AT000	2	CAN000		HAVE00	
WITH00	17	ONE00		BUT000	1	ITS000	4
ON0000	1	THIS0	4	WHAT00		THEY00	
TWO000		FOR00	8	HAS000	3	THERE0	2
NEW000		ABOUT	1	ONLY00	1	THEIR0	
EACH00		MORE0		WILL00		IF0000	
INTO00	1	MUCH0		OR0000	1	OTHER0	1
SAME00	6	SOME0	1	WAS000	19	WHEN00	5
ALL000		ALMOST		BEEN00	2	MOST00	
SEE000		THESE	1	ANY000		ITSELF	
LEAST0		MANY0		NOT000		OUT000	1
MAY000	1	SO000		THAN00		THEN00	1
THREE0	2	TO000		VERY00		PERHAP	
ACROSS		ALSO0	1	HIGHER		LIKE00	
SUCH00		WAY00		ABLE00		ABOVE0	4
AFTER0	4	AGAIN		AGAINS		AGO000	
ALLOWS		ALONG		ALREAD		ALTHOU	1
AMONG0		ANOTHE		APPEAR	2	APPARE	
ARISE0		AROUND		AWAY00		BACK00	
BECAME		BECAUS		BECOME		BEFORE	
BEING0	1	BELOW		BESIDE		BEST00	
BETTER		BETWEE		BOTH00		BROUGH	
CALLED		CAME0	1	CANNOT		CENT00	
CERTAI		CLEARL		COME00		COMPLE	
EARLY0		EASY0		COULD0	1	DEPEND	
DID000		DO000		DOES00		DOUBTL	
DOWN00		DUE00		DURING		EARLIE	
EASILY		EITHER		ENOUGH		ESPECI	
EVEN00		EVENTU		EVERY0		FAR000	
FEW000		FINALL		FIND00		FIRST0	1
FOUND0		FURTHE	3	GAVE00		GET000	
GIVE00		GIVEN	1	GIVES0		GOES00	
GOING0		GREAT		GREATE		HAD000	3
HAPPEN		HE000		HER000		HERE00	1
HIGH00		HIM00		HIMSEL		HIS000	
HOW000		HOWEVE	1	I00000		INCLUD	
INDEED		INSTEA		JUST00		LARGE0	
LARGEL		LAST0		LATER0		LEFT00	
LIKELY	1	LITTLE	1	LONG00		LOW000	
MADE00		MAKE0		MAKES0		ME0000	
MEANS0		MERELY		MIGHT0		MOREOV	
MUST00		MY000		NEARLY		NEED00	
NEEDED		NEEDS		NEXT00		NO0000	3
NONE00		NOW00		OFF000		OFTEN0	
ONCE00		OTHERS		OUR000		OVER00	2
OWN000		PARTLY		SHOWS0		SINCE0	
SMALL0	2	SOMETH		SOMETI		SPECIA	
PER000	2	POSITI		PRESEN	3	QUITE0	
RATHER		READIL		REALLY		REMAIN	
RIGHT0		SAID0		SECOND	1	SEEM00	
SEEMS0		SEEN0		SERVES		SEVERA	
SHE000		SHOULD		SHOW00		SHOWN0	2
STILL0	1	TAKE0		TAKEN0	1	TAKING	
THEM00		THEMSE		THEREF		THINGS	
THIRD0		THOSE		THOUGH	1	THROUG	
THUS00		TIMES	1	TOOK00		TOGETH	
TOWARD		UNDER	2	UNTIL0		UP0000	1
UPON00		US000		USED00	1	USUALL	
VARIOU		WHILE		WERE00	7	WENT00	
WAYS00		WELL0	1	WHOLE0	0	WHO000	
WHERE0		WHOM0		WHOSE0	0	WHY000	
WITHOU		WOULD		YET000	2	0CW400	25

```
1DOC  R283
TABLE OF SENTENCE LENGTHS IN WORDS AND NUMBER OF SENTENCES EACH

 SENTS
  1    0  2    0  3    0  4    0  5    0  6    0  7    0  8    0  9    0  10    1
 11    0 12    0 13    2 14    1 15    0 16    0 17    4 18    1 19    0  20    4
 21    4 22    2 23    2 24    2 25    1 26    1 27    2 28    1 29    2  30    1
 31    0 32    1 33    0 34    0 35    2 36    0 37    0 38    0 39    0  40    0
 41    0 42    0 43    0 44    0 45    1 46    0 47    0 48    0 49    0  50    0
 51    0 52    0 53    0 54    0 55    0 56    0 57    0 58    0 59    0  60    0
 61    0 62    0 63    0 64    0 65    0 66    0 67    0 68    0 69    0  70    0
 71    0 72    0 73    0 74    0 75    0 76    0 77    0 78    0 79    0  80    0
 81    0 82    0 83    0 84    0 85    0 86    0 87    0 88    0 89    0  90    0
 91    0 92    0 93    0 94    0 95    0 96    0 97    0 98    0 99    0 100    0

NUMBER OF NON-COMMON WORDS HAVING FREQUENCIES 1 TO 100
1DOC  R283

 FR WDS FR WDS FR WDS FR WDS FR WDS FR WDS FR WDS FR WDS FR WDS  FR WDS
  1 134  2  45  3  18  4   7  5   7  6   3  7   1  8   0  9   1  10   0
 11   0 12   1 13   0 14   0 15   0 16   2 17   0 18   0 19   0  20   0
 21   0 22   0 23   0 24   0 25   0 26   0 27   0 28   0 29   0  30   0
 31   0 32   0 33   0 34   0 35   0 36   0 37   0 38   0 39   0  40   0
 41   0 42   0 43   0 44   0 45   0 46   0 47   0 48   0 49   0  50   0
 51   0 52   0 53   0 54   0 55   0 56   0 57   0 58   0 59   0  60   0
 61   0 62   0 63   0 64   0 65   0 66   0 67   0 68   0 69   0  70   0
 71   0 72   0 73   0 74   0 75   0 76   0 77   0 78   0 79   0  80   0
 81   0 82   0 83   0 84   0 85   0 86   0 87   0 88   0 89   0  90   0
 91   0 92   0 93   0 94   0 95   0 96   0 97   0 98   0 99   0 100   0

TABLE SHOWING PERCENTAGE OF OCCURRENCES OF NON-COMMON WORDS
AND PERCENTAGE OF DIFFERENT NON-COMMON WORDS HAVING FREQUENCIES 1 TO 12

 FRQ PCTNCO PCTDIF FRQ PCTNCO PCTDIF FRQ PCTNCO PCTDIF FRQ PCTNCO PCTDIF
   1 .31981 .61187   2 .21480 .20548   3 .12868 .08219   4 .06683 .03196
   5 .08353 .03196   6 .04296 .01370   7 .01671 .00457   8 .00000 .00000
   9 .02148 .00457  10 .00000 .00000  11 .00000 .00000  12 .02864 .00457

TABLES OF GROUPINGS OF WORDS, SHOWING FOR EACH GROUP NUMBER OF
OCCURRENCES, NUMBER OF DIFFERENT WORDS, NUMBER OF WORDS PER SENTENCE,
AVERAGE FREQUENCY, PERCENTAGE OF ALL DIFFERENT WORDS AND PERCENTAGE
OF ALL OCCURRENCES
 TYP OUTPUT OCCUR DIF WDS WD PER SENT AVG FREQ   PCT DIFFRNT PCT OCCURNC

 TOTAL WORDS   806   291     23.0286      2.7698
 COMMN WORDS   387    72     11.0571      5.3750    0.247423     0.480149

TABLE OF LENGTHS OF NON-COMMON WORDS BY NUMBER OF LETTERS

 WRD LENG  1     0     0     0.0000      0.0000    0.000000     0.000000
 WRD LENG  2     0     0     0.0000      0.0000    0.000000     0.000000
 WRD LENG  3    13     6     0.3714      2.1667    0.020619     0.016129
 WRD LENG  4    53    32     1.5143      1.6562    0.109966     0.065757
 WRD LENG  5    45    26     1.2857      1.7308    0.089347     0.055831
 WRD LENG  6    38    25     1.0857      1.5200    0.085911     0.047146
 WRD LENG  7    57    35     1.6286      1.6286    0.120275     0.070720
 WRD LENG  8    40    28     1.1429      1.4286    0.096220     0.049628
 WRD LENG  9    51    36     1.4571      1.4167    0.123711     0.063275
 WRD LENG 10    32    19     0.9143      1.6842    0.065292     0.039702
 WRD LENG 11    20    15     0.5714      1.3333    0.051546     0.024814
 WRD LENG 12    13     6     0.3714      2.1667    0.020619     0.016129
 WRD LENG 13    10     5     0.2857      2.0000    0.017182     0.012407
 WRD LENG 14    13     4     0.3714      3.2500    0.013746     0.016129
 WRD LENG 15    20     4     0.5714      5.0000    0.013746     0.024814
 WRD LENG 16     1     1     0.0286      1.0000    0.003436     0.001241
 WRD LENG 17     3     3     0.0857      1.0000    0.010309     0.003722
 WRD LENG 18    10     4     0.2857      2.5000    0.013746     0.012407

TABLE OF LENGTHS OF COMMON WORDS BY NUMBER OF LETTERS

 CWD LENG  1     6     1     0.1714      6.0000    0.003436     0.007444
 CWD LENG  2   137    15     3.9143      9.1333    0.051546     0.169975
 CWD LENG  3   138    12     3.9429     11.5000    0.041237     0.171216
 CWD LENG  4    64    18     1.8286      3.5556    0.061856     0.079404
 CWD LENG  5    28    17     0.8000      1.6471    0.058419     0.034739
 CWD LENG  6     6     5     0.1714      1.2000    0.017182     0.007444
 CWD LENG  7     7     3     0.2000      2.3333    0.010309     0.008685
 CWD LENG  8     1     1     0.0286      1.0000    0.003436     0.001241
 CWD LENG  9     0     0     0.0000      0.0000    0.000000     0.000000
 CWD LENG 10     0     0     0.0000      0.0000    0.000000     0.000000
```

```
TYP OUTPUT OCCUR DIF WDS WD PER SENT AVG FREQ   PCT DIFFRNT PCT OCCURNC

TABLE OF WORDS HAVING FREQUENCIES IN RANGES INDICATED -100 TO 91,
90 TO 81, ETC.

GRP -       91     0      0        0.0000        0.0000      0.000000      0.000000
GRP -       81     0      0        0.0000        0.0000      0.000000      0.000000
GRP -       71     0      0        0.0000        0.0000      0.000000      0.000000
GRP -       61     0      0        0.0000        0.0000      0.000000      0.000000
GRP -       51     0      0        0.0000        0.0000      0.000000      0.000000
GRP -       41     0      0        0.0000        0.0000      0.000000      0.000000
GRP -       31     0      0        0.0000        0.0000      0.000000      0.000000
GRP -       21     0      0        0.0000        0.0000      0.000000      0.000000
GRP -       11    44      3        1.2571       14.6667      0.010309      0.054591
GRP -        1   375    216       10.7143        1.7361      0.742268      0.465261


TABLE OF FREQUENCY GROUPS BY TENTHS OF NON-COMMON OCCURRENCES

FREQ        12    44      3        1.2571       14.6667      0.010309      0.054591
FREQ         5   113     15        3.2286        7.5333      0.051546      0.140199
FREQ         4   141     22        4.0286        6.4091      0.075601      0.174938
FREQ         3   195     40        5.5714        4.8750      0.137457      0.241935
FREQ         2   285     85        8.1429        3.3529      0.292096      0.353598
FREQ         1   419    219       11.9714        1.9132      0.752577      0.519851


TABLE OF COMMON WORDS BY INITIAL LETTER

COMM INIT A     57     11          1.6286        5.1818      0.037801      0.070720
COMM INIT B     13      5          0.3714        2.6000      0.017182      0.016129
COMM INIT C      2      2          0.0571        1.0000      0.006873      0.002481
COMM INIT F     16      4          0.4571        4.0000      0.013746      0.019851
COMM INIT G      1      1          0.0286        1.0000      0.003436      0.001241
COMM INIT H      8      4          0.2286        2.0000      0.013746      0.009926
COMM INIT I     36      5          1.0286        7.2000      0.017182      0.044665
COMM INIT L      2      2          0.0571        1.0000      0.006873      0.002481
COMM INIT M      1      1          0.0286        1.0000      0.003436      0.001241
COMM INIT N      3      1          0.0857        3.0000      0.003436      0.003722
COMM INIT O     43      7          1.2286        6.1429      0.024055      0.053350
COMM INIT P      5      2          0.1429        2.5000      0.006873      0.006203
COMM INIT S     13      6          0.3714        2.1667      0.020619      0.016129
COMM INIT T    107     11          3.0571        9.7273      0.037801      0.132754
COMM INIT U      4      3          0.1143        1.3333      0.010309      0.004963
COMM INIT W     49      5          1.4000        9.8000      0.017182      0.060794
COMM INIT Y      2      1          0.0571        2.0000      0.003436      0.002481


TABLE OF NON-COMMON WORDS BY INITIAL LETTER

NCOM INIT 0      1      1          0.0286        1.0000      0.003436      0.001241
NCOM INIT 1      5      3          0.1429        1.6667      0.010309      0.006203
NCOM INIT 2      2      2          0.0571        1.0000      0.006873      0.002481
NCOM INIT 3      4      1          0.1143        4.0000      0.003436      0.004963
NCOM INIT 6      1      1          0.0286        1.0000      0.003436      0.001241
NCOM INIT A     30     18          0.8571        1.6667      0.061856      0.037221
NCOM INIT B     29      9          0.8286        3.2222      0.030928      0.035980
NCOM INIT C     27     15          0.7714        1.8000      0.051546      0.033499
NCOM INIT D     22     14          0.6286        1.5714      0.048110      0.027295
NCOM INIT E     20     10          0.5714        2.0000      0.034364      0.024814
NCOM INIT F     13      7          0.3714        1.8571      0.024055      0.016129
NCOM INIT G      1      1          0.0286        1.0000      0.003436      0.001241
NCOM INIT H      8      6          0.2286        1.3333      0.020619      0.009926
NCOM INIT I     25     15          0.7143        1.6667      0.051546      0.031017
NCOM INIT J      2      1          0.0571        2.0000      0.003436      0.002481
NCOM INIT K      3      2          0.0857        1.5000      0.006873      0.003722
NCOM INIT L      8      7          0.2286        1.1429      0.024055      0.009926
NCOM INIT M     32     12          0.9143        2.6667      0.041237.     0.039702
NCOM INIT N     37      9          1.0571        4.1111      0.030928      0.045906
NCOM INIT O     26     16          0.7429        1.6250      0.054983      0.032258
NCOM INIT P     22     15          0.6286        1.4667      0.051546      0.027295
NCOM INIT R     38     20          1.0857        1.9000      0.068729      0.047146
NCOM INIT S     35     18          1.0000        1.9444      0.061856      0.043424
NCOM INIT T     11      7          0.3143        1.5714      0.024055      0.013648
NCOM INIT U      3      2          0.0857        1.5000      0.006873      0.003722
NCOM INIT V      9      3          0.2571        3.0000      0.010309      0.011166
NCOM INIT W      4      3          0.1143        1.3333      0.010309      0.004963
NCOM INIT Y      1      1          0.0286        1.0000      0.003436      0.001241
```

Conclusion

The potentialities of auto-encoding of documents from machine-readable texts have been brought forth by way of examples typical of the Information Retrieval research work presently in progress at the IBM Research Division. Some of the processes discussed are being tested through pilot operations within and outside of the IBM Corporation. While the feasibility of these processes has been established in principle, their effectiveness with respect to the human user might not be satisfactorily established until a system has been in full size operation over a considerable period of time.

Amongst the difficulties encountered in the processing of machine readable texts, inconsistencies in the use of punctuation marks, compounds, capitals, spacing and indentations have been a problem way out of proportion with respect to the simple functions these devices stand for. For instance, even with the aid of a dozen different tests performed by the machine, the true end of a sentence cannot be determined with certainty. It is hoped that publishers of scientific literature will in time sacrifice some of the niceties and aesthetic aspects of the printed page for the sake of clarity in communication.

H. P. Luhn
May 15, 1959
L # 435