The last ten years of equipment development and the application of mechanization techniques are reviewed for each of several functionally separate approaches, such as punched-card systems, computer systems, and magnetic media systems. Comments are made on the progress to date and the degree of activity to be expected during the next few years for each of these approaches.

The Historical Development and Present State-of-the-Art of Mechanized Information Retrieval Systems

CHARLES P. BOURNE

Stanford Research Institute, Menlo Park, Calif.

• Introduction

Within the last ten years, and especially within the last two or three, there has been an increasing interest in techniques and tools to fully or partially mechanize some of the operations of information storage and retrieval systems. The interest has been shared by equipment manufacturers, information custodians, and the information generators and users. The development of the mechanization techniques has come primarily as a result of:

1) The support of a rather large research effort in this field by military agencies in order to obtain solutions to particular problems of interest to them. (Fortunately a great deal of this research is also applicable to many general documentation problems.)
2) A continuous long-range research program by several commercial organizations in the expectation of developing this particular market.
3) The recent support by agencies such as the National Science Foundation and the Council on Library Resources for research on problems of more general interest to the documentation field.
4) The curiosity and academic interest of a small number of individuals who have a basic training in fields other than documentation.

As a side comment, it might be noted that the greatest amount of support for research on the mechanization of information retrieval has been provided by the U. S. Air Force.

This increased research and development effort, during the last few years, represents an increasing awareness and realization of the problems. It also represents the natural extension of conventional techniques and equipment to more unusual problems. For example, punched-card equipment, computers, and microfilm systems were seriously directed toward information retrieval applications only after this equipment had demonstrated its effectiveness in many stereotyped commercial applications.

• A Look at the Developments

Edge-punched cards. There has been a tremendous increase in the use of edge-punched cards during the last ten years. However, applications which are closely related to the classical documentation problems (e.g., compilation of indexes, bibliographies, charge systems, and fact files) form a very small portion of the commercial market for these cards. The Zator card is the only edge-punched card that was designed and used specifically for documentation purposes. It is expected that there will soon be further development work in this area, although there have not been any basically new developments in this subject for several years. The edge-punched cards will continue to be used in increasing numbers.

Punched cards and tabulating equipment. A great many individuals and organizations have used punched tabulating cards and electronic accounting machines (EAM) for storage and retrieval applications as well as other related tasks such as reproduction and dissemination. In most instances, the applications were implemented with EAM equipment that was available but used primarily for some other purpose such as general accounting. Only a few organizations are able to justify the acquisition of this equipment solely for storage and retrieval applications. This situation similarly exists in the application of computer equipment to documentation activities.

Although all of the card systems reportedly used general purpose EAM equipment to good advantage, several new units, all developed by International Business Machines (IBM), are:

Reprinted from AMERICAN DOCUMENTATION, Vol. 12, No. 2, April, 1961

IBM-9900 Special Index Analyzer (a commercial modification of the COMAC, which was developed by Documentation, Inc., under U. S. Air Force sponsorship)
IBM-9310 Universal Card Scanner
IBM-101 with Row-by-Row Scanning Attachment (a modification to the existing IBM-101 Statistical Sorter)

For a number of reasons, primarily economic, very little use has been made of these special systems, and under the present circumstances it is unlikely that these particular units will receive any substantial utilization in the immediate future. However, the use of the standard EAM equipment will continue to rise, particularly in some of the newer applications. The use of card equipment for the preparation and composition of information for printing (e.g., library catalogs and the earlier *Index Medicus*, etc.) is an example of an application which will become a great deal more common in the years to come.

Computer equipment. As computer equipment has become more readily available, more organizations and individuals have programmed them for storage and retrieval applications. With very few exceptions, the information retrieval programs have represented a very small fraction of the total workload of the respective installations. Some variation of a retrieval program, usually for a collection of technical reports and documents, has been demonstrated on nearly every major type of computer in operation today. With the exception of a few complex programs for working with patents and chemical compounds, the programs have usually represented some form of coordinate indexing.

In addition to file searching, the computers have also been used for the generation of conventional catalogs and indexes to be used in manual systems; the automatic routing or dissemination of material to potentially interested parties; the preparation of abstracts from straight textual material; the generation of lists of key index words; and the generation of permutation or keyword-in-context indexes. Good or bad, and regardless of the intellectual arguments which accompany them, these applications and techniques are here to stay, and will continue to be used by an increasing number of organizations.

Essentially all of this original work with computers has been done by the government and commercial organizations who are technical information users and distributors, rather than the universities and library schools. To some degree, this may be due to the unavailability of computer equipment, although it must be noted that there are more than 100 computer systems currently operating in U. S. universities and colleges.

Magnetic media. Several magnetic tape and card systems have been developed specifically for file-searching applications. The systems which have been demonstrated to date are:

Logic Processor (Aeronutronics)
Index Searcher (Computer Control Co., Inc.)
Univac Tape Searchwriter (Remington Rand)
Findafact (Rese Engineering Co.)
GE-250 Information Seaching Selector (General Electric Co.)
Magnacard (Magnavox Co.)
Tape Searcher (Herner & Co.)

Only one or two developmental models of each of the above units have been produced to date, and none of them is completely operational yet in a documentation system. Several other special data processing systems have been developed which perform tape searching as an auxiliary operation.

It should be noted that the functions of a tape-searching device may also be performed just as well on a computer system, and the recent commercial availability of several moderately priced computer systems has apparently priced most of the present tape-searching systems (approximately $100,000 each) out of the market. With the exception of a few very special systems, it is unlikely that very many tape searchers (with the same price and capability as the systems noted above) will be used in the next few years. The relatively inexpensive systems, such as Herner's Tape Searcher (approximately $10,000) should find a more receptive market. The more complex tape-searching equipment will probably be used primarily by special information centers (e.g., the operational Western Reserve University-American Society for Metals system) and a few organizations with very large and special file problems.

Image-storage systems. The development of mechanized image-storage and retrieval systems will continue to be one of the largest areas of activity. To date, the Filmorex and Minicard equipment are the only systems which can be considered to be completely operational. And with the exception of a single Minicard test system at Eastman Kodak, all of this currently operating equipment in the United States (i.e., a single Filmorex system and four Minicard systems) is being used by government installations.

Additional systems and system components which are still under development, or have been developed but not actually used for documentation activities, are listed below:

Rapid Selector (National Bureau of Standards)
Finder-Reader System (M.I.T.)
Walnut (IBM Corp.)
Film Searcher (Rabinow Engineering Co., Inc.)
Verac 903 (AVCO Corp.)
Filesearch (FMA, Inc.)
FLIP (Benson-Lehner Corp.)
Itek-card (Itek Corp.)
Magnacard with image (Magnavox Co.)
Film Searcher (Magnavox Co.)
Telecard (General Precision Laboratories, Inc.)
Rapid Access Look-up System (Ferranti-Packard).

There is a wide range of needs for complete image systems, ranging from the requirements of the individual researcher to the requirements of large and specialized file collections. As with many of the other mechanized schemes, the cost factors have been high enough to discourage most of the potential users. However, it is expected that within the next year or two, several image systems will be developed and marketed which offer economic as well as operational advantages.

It is expected that there will continue to be an increased utilization of unit-record microfilm systems, with particular emphasis on forms such as the microfilm aperture card, continuous-image card, and micro-opaque cards. The increased popularity of these systems is due in part to the recent improvements in microfilm viewers, copiers, and printers.

Character-recognition equipment. Character- and page-reading equipment has been developed to such a point that there are presently at least 150 machines in operation for reading characters one line at a time (e.g., bank checks, retail charge slips, etc.), and several units in operation for reading entire pages of typewritten English text (e.g., as input devices for data processing or communications systems). In its present form, the character-sensing equipment will not be applicable to documentation problems for several years to come. Likewise, the page-reading equipment will be severely limited in its use. This is primarily because of (1) the restrictions of the techniques (e.g., the rather strict requirements on the type, quality, and format of the printing), as well as (2) the relatively little need for this type of capability in equipment for general documentation systems. However, there are many special operations (e.g., direct text scanning or searching, automatic preparation of indexes, correlation and organization of text material, etc.) which could profitably use page-reading equipment.

• Summary and Conclusions

Compared to the earlier years, there has been a very great effort during the last decade to develop and use mechanized storage and retrieval systems. This activity will continue at least at the same level of effort, with the majority of effort spent on the development of image systems and special digital storage systems, as well as new programming efforts for the utilization of general-purpose digital computers.

To date, the majority of the mechanized systems have adopted classification and indexing techniques which were developed primarily for manual systems. In the years to come, there will be additional efforts to develop and improve techniques which offer special advantages for machine operation, such as prime-number coding, superimposed coding, and abbreviated spelling techniques. In any case, a good classification or indexing system will continue to be essential to the proper operation of any mechanized system.

There is every reason to believe that a human element must remain an integral part of the operating system, and that there will certainly continue to be at least as much need for trained information specialists and documentalists as there is today.

The actual design and development of the mechanized systems will continue to come primarily as a result of interdisciplinary efforts, and it should be obvious that this situation can be worked to the best advantage of all parties concerned only if the members of the various disciplines make a more earnest effort to educate, to cooperate, and to contribute to the joint solution of these problems.

Bibliography

1. United States Senate. May 24, 1960. 86th Congress, 2nd Session. Report of the Committee on Government Operations: Documentation, indexing, and retrieval of scientific information.
2. National Science Foundation, Office of Scientific Information, Washington, D. C. Current research and development in scientific documentation. Nos. 1-6, July 1957-May 1960; Non-conventional technical information systems in current use. No. 1 (Jan. 1958), No. 2 (Sept. 1959), Supp. No. 2 (March 1960).
3. Stanford Research Institute, Menlo Park, Calif. Bibliographies by Charles P. Bourne. Bibliography on the mechanization of information retrieval (Feb. 1958), 22 pp.; Supp. I (Feb. 1959), 25 pp.; Supp. II (Feb. 1960), 14 pp.; Supp. III (Feb. 1961), 27 pp.

A STUDY OF METHODS FOR SYSTEMATICALLY ABBREVIATING ENGLISH WORDS AND NAMES

BY

CHARLES P. BOURNE

AND

DONALD F. FORD

STANFORD RESEARCH INSTITUTE

MENLO PARK, CALIFORNIA

Reprinted from JOURNAL OF THE ASSOCIATION FOR COMPUTING MACHINERY
Volume 8, Number 4, October 1961
Printed in U.S.A.

A Study of Methods for Systematically Abbreviating English Words and Names*

CHARLES P. BOURNE AND DONALD F. FORD

Stanford Research Institute, Menlo Park, California

Abstract. This study investigated various techniques for systematically abbreviating English words and names. Most of the attention was given to the techniques which could be mechanized with a digital device such as a general purpose digital computer. Particular attention was paid to techniques that could process incoming information without prior knowledge of its existence (i.e., no table lookups). Thirteen basic techniques and their modifications are described. In addition, most of the techniques were tested on a sample of several thousand subject words and several thousand proper names in order to provide a quantitative measure of comparison.

Introduction

There are many instances in which it may be advantageous to abbreviate[1] English words, such as people's names, places, street addresses, proper nouns, or continuous text material. Common subjects of abbreviation techniques are addresses in telephone directories, customer account number identifications in some data processing systems, and commercial teletype or ham radio vocabularies. There will always be a need for abbreviation or coding schemes in order to efficiently utilize the storage media in computers, punched card systems, and other storage and processing systems. Abbreviation techniques may also be applied to improve the efficiency of some communication systems by decreasing the number of characters which must be transmitted. If the abbreviations are to be made automatically or semi-automatically, they must be obtained by some systematic method.

This study concentrated on techniques which could be mechanized with a digital device such as a general purpose computer. Some of the abbreviation schemes (i.e., selective drop-out by letter position, by separate character frequency distributions for each letter position, by bigram rankings, by arithmetic methods, and shuffling) appear to be new and different and in some cases provide improved performance over the remaining techniques.

Most of the schemes described in this paper were empirically tested with a computer, using a rather large and representative sample data base. In this manner, some quantitative comparisons of the merits of the various methods

* Received January, 1961.

[1] In this paper we often refer to the "abbreviation" of words, but actually we are concerned with "coding" the words, or transforming the source word into some different pattern. In just happens that in satisfying some of the objectives of our code design we actually "abbreviate" the word. This comment is made because some of the schemes and transformations proposed in this paper bear little resemblance to what we normally consider as "abbreviations."

could be made. The criteria for comparison and evaluation of the systems are described in subsequent sections.

The authors were primarily interested in the abbreviation of subject words, such as library catalog subject headings or descriptors that were used with documentation systems. The sample data base, the supporting statistical data, and the experiments were primarily oriented toward this problem. However, additional attention was paid to the abbreviation of people's names. No effort was made to investigate word groups, such as word pairs or continuous text. However, it does appear that some very interesting work can be done in this area.

General Objectives

The intent of this study was to find methods to systematically code English words while satisfying the following general objectives:

(1) Each word should be coded to require as little storage space as possible.

(2) Each word should retain the same degree of discrimination and uniqueness that it had in the original sample. (For example, if there were 2082 unique words in the original sample, then hopefully there would still be 2082 unique items after abbreviation.)

(3) If possible, each word should retain some mnemonic similarity to the original word. (For this reason, the initial letter was automatically retained with all the schemes tested except for the technique which truncated the left end of the word.)

(4) The procedure should not rely on any prior knowledge of the population of words which must be abbreviated. (For example, one of the most efficient coding schemes would be to construct a table with code numbers for each word that would be encountered. But this would require the development of, and continued reference to, a table.) It was the intent of this study to work on schemes which could generally accept and abbreviate any word that was presented to the system, without requiring the use of any tables that were generated in anticipation of the word.

(5) It would be a useful feature if the abbreviated word could be systematically transformed back to the original word when desired.

Criteria for Comparative Evaluations

It was assumed that the major objective was to provide as much condensation as possible while maintaining a maximum amount of discrimination between the sample words. The degree of success in meeting this particular objective can be measured more quantitatively than it can be for any of the other objectives. In particular, the degree of success for each technique was measured by using each particular technique to abbreviate the test material to a given number of characters, and then counting the number of unique entries which remained after the abbreviation process. Each scheme was tested, and the degree of uniqueness was measured for various amounts of abbreviation. These experiments resulted in the

TABLE I. *Number of unique codes possible with a series of alphabetic letters*

Number of characters used in the abbreviation	Number of unique codes possible
1	26
2	676
3	17,576
4	456,976
5	11,821,376
N	26^N

curves which are shown in Figures 1, 2, and 3, and are described later in this paper. These curves provide some measure of the performance of each particular scheme. In this paper, the term "operating characteristics" was used to describe the results of these empirical studies. That is, the figures give a summary of the "operating characteristics" of the various schemes reported in this study.

As illustrated by Table I, the 26 characters of the English alphabet offer a large number of coding possibilities, so that a few alphabetic characters are capable of providing a relatively large degree of discrimination.

Data Base

The primary source of English words for the test material was a list of 2082 different single-word descriptors (e.g., "magnetic," "optical") which were used to index a collection of technical documents at Stanford Research Institute. This sample did not represent the entire descriptor dictionary, since multiple word descriptors (e.g., "aerodynamic heating" or "black body") were not considered for these tests. In all cases the tests started with the full and correct spelling of the words. That is, no truncated or abbreviated words were used in the data base.

The second source of test material was a list which represented the entire 1959 student registration of Stanford University. Each of the names was initially restricted to a field of 25 characters and consisted of at least the last name, as well as some combination of the first and middle (or more) names, and their prefixes. The names were basically more awkward to work with than the subject words, and a simplification was made by editing each name to remove all spaces and special characters (e.g., hyphens) and compressing it to form a single long word of 22 or less characters. The original name list contained 8207 names. When the duplicates were removed, 8184 unique names remained for the data base.

Description of the Various Coding Methods

Most of the following methods were tested with the word file and some of the methods were tested in the name file. In nearly all cases the initial letter was retained and the scheme was not executed if the word length was initially less than or equal to the desired word length.

Simple Truncation of the Right End. Starting from the right end of the word, drop off letters until the required word size is obtained. No other operations or exceptions are allowed.

Simple Truncation of the Left End. This is the same process as above, except that the process drops off the letters from the left end of the word.

Elimination of Vowels. Starting from the right end of the word eliminate vowels (a, e, i, o, u) until the desired word length is reached. If the word cannot be sufficiently shortened by this method, then use simple truncation of the right end to reach the desired length after the available vowels are eliminated.

Selective Drop-out by Letter Position. Starting from the left end of the word, eliminate every nth letter. That is, eliminate every 2nd letter, or every 3rd letter, or every 4th letter. (Each different n constitutes a different technique to be processed and tested separately.) An illustration of this technique to reduce a word to three characters by dropping out every 2nd character (i.e., $n = 2$), is given by the following two examples (letters to be dropped are shown in bold face.)

1st pass.....	**A**B**L**A**T**I**O**N	**A**C**C**E**P**T**A**N**C**E
2nd pass.....	**A**L**T**O	**A**C**P**A**C**
Final result.....	A**T**O	A**P**C

Selective Drop-out by a Single Ranking of Letter Usage. Given an empirical ranking of the composite[2] frequency of usage of the 26 alphabetic characters, eliminate the most common letters until the desired word length has been achieved. In case of a tie choice, remove the right-most letters first. This technique was originally suggested by Luhn [14].

The following composite letter rankings were used in this test, and were derived separately from each of the two different data samples.

SUBJECT WORDS

E I R O A T N S L C P M D U H G Y B F V K W X Z J Q

NAMES

E A R N L O I S T H D M C B G U W Y J K P F V Z X Q

Other frequency distributions which have been published, or suggested for this application were not used, because they did not accurately represent the statistical nature of the sample data. This is probably because the majority of published frequency distributions were obtained by examining continuous text material which is biased by the presence of many common words (e.g., the, of, and, in, a). A more detailed study of the statistical nature of English words, and a comparison of the rankings obtained in this study with some of the previously published rankings has been made in a companion paper [38].

[2] That is, an average frequency of usage, which does not consider that the letters might be used differently in the various letter positions. Actually, there are some distinct differences in the character frequency distributions for the various letter positions.

Selective Drop-out by Separate Rankings of Character Usage for each Letter Position. Given a separate empirical ranking of letter usage for *each* letter position, examine each word to determine the respective ranking of each of the characters present. Then remove the letters in the order of their popularity until the desired word length has been obtained. That is, remove the most popular letters first. In case of a tie choice, remove the right-most letters first.

It has been found that there is a remarkable similarity in the ranking and distribution of most of the characters, regardless of their letter position. However, for both subject words and proper names the distributions of the first two, and possibly three, letter positions are each markedly different from the distributions of all the other letter positions. In particular, the popularity of vowels in the second letter position suggests that the second letter position does not provide a great deal of discrimination.[3]

Selective Drop-out by a Single Ranking of Bigram (Letter Pair) Usage. Given an empirical ranking of the composite frequency of occurrence of bigrams (i.e., ignoring the position of bigram occurrence within the words), examine each word to determine the ranking of all the possible letter pairs (adjacent letters only). The elimination process considers the association of each letter with the neighboring letters, and rejects the letters which make up the most common bigrams. With the exception of the initial letter (which was retained by policy), every letter contributes to two different bigrams. The total ranking of the popularities of those two bigrams determines the selection criteria for that particular letter. Each of the letters is evaluated in this way to derive an index or measure of discrimination for that particular letter. The least descriminating letters are removed to arrive at the desired word length. An illustration of this process is given by the following example, reducing the word ACETYLENE to four characters:

Bigram rankings....	66	58	25	108	164	24	20	26	2	
	∧	∧	∧	∧	∧	∧	∧	∧	∧	
Letter.....	A	C	E	T	Y	L	E	N	E	(Space)
Letter index.....		124	83	133	272	188	44	46	28	
Final result.....	A	T	Y	L						

It would also be possible to operate with trigrams or larger groups of letters, and consideration could be given to letter pairs which are separated by one or more letters. It would also be interesting to look at the use of a separate bigram ranking for each letter position.

The bigram rankings and distributions for this sample data as well as a discussion of other previously published bigram distributions has been given in a companion paper [38].

Truncation after Elimination of the Second Character. Automatically eliminate the second character of the word, counting from the left side of the word. Then

[3] This has been recognized by other workers, notably Cox, Casey, and Bailey [7], who have designed punched card codes for coding proper names which either ignored the second letter, or provided separate coding spaces for each vowel, and one or two additional code spaces to take care of the remaining 21 characters.

perform simple truncation, starting at the right end of the word. No other operations or exceptions are allowed. This process is a modification of the simple truncation scheme, but seeks an improvement by omitting the second letter with its weak discriminating power.

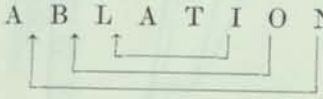
The Use of a Check Digit (or Letter). The discriminating powers of any proposed scheme will probably be enhanced by using a check digit[4] or check letter to describe the letters which have been eliminated. For example, an abbreviation technique may reduce COMPUTERS and COMPUTATION to the same form COMPUT, with a resulting loss in discrimination. Using a check letter, the original words would abbreviate to COMPU*, with the letter in the asterisk position generally being different for each original word. There are several ways to generate a check letter, such as (1) add up the numbers which represent the position of the eliminated letters in the alphabet (e.g., A=1, B=2, $\cdots$, Z=26), or (2) add up the numbers which represent the frequency rankings of the eliminated letters (e.g., E=1, A=2, $\cdots$, Q=26 for names), or (3) count the number of eliminated letters. If the number resulting from any of these three operations is larger than 26, then cast out 26 as many times as possible, with the remainder (less than 26) to be used as the entry to a table of check letters. The table of check letters could be a simple listing of the alphabet. An illustration of one method for deriving a check digit is given by the following example in which the word ABLATION is reduced to a total of four characters by simple truncation and the addition of a check letter:

	1 20 9 15 14 = 59 = 26 + 26 + 7 = H
Original word.....	A B L **A T I O N**
Final abbreviation.....	A B L H

(Note: the table for the eliminated letters uses A=1, $\cdots$, Z=26; but the table for the derivation of the check letters uses values A=0, $\cdots$, Z=25 since it is possible to obtain the value zero if the first number exactly equals a multiple of 26.)

In this study the check digit technique was used to enhance two different methods: (1) simple truncation of the right end, and (2) selective drop-out of every second letter (i.e., $n=2$), after they had been run without check letters.

Shuffle and Truncation. Shuffle the letters of the word and then drop the letters from the right end of the word until the desired word length is obtained. The shuffle consists of a simple folding of the letters within the word, as illustrated by the following example in which the word ABLATION is reduced to four letters. Different types of folding patterns can also be used.



Elimination of Miscellaneous Redundancies. Some benefit may be achieved by

[4] Check digits are generated and used in many ways with data processing systems. However, the application of check digits to word abbreviation was first suggested by Luhn [14].

such tricks as the deletion of U after Q, or one of the double letters when they occur together (e.g., tt, mm, nn), or the second vowel of a double vowel (e.g., ae, io, oe). These variations are not frequent enough to be used alone as a complete technique, but it is possible that they may enhance other techniques.

Arithmetic Manipulations. On some types of computer equipment it is possible to do such things as square the original word, and retain some of the digits of the result for use as the abbreviated word. There would no longer be a resemblance to the original word, but the transformation might yield a relatively unique expression. These expressions would probably be represented as numbers, or mixed alphanumeric and special characters. Many other types of arithmetic operations could also be used.

Soundex Code. The conventional Soundex system for manual file operations converts names to a code word of one alphabetic character and 3 numeric characters. It is possible to generate larger code words, but this is seldom done. The conversion rules are:

(1) Always retain the first letter of each name (counting from the left end).
(2) Drop out A, E, I, O, U, Y, W, and H.
(3) Assign the following numbers to the remaining similar-sounding sets of letters:

B, F, P, V = 1
C, G, J, K, Q, S, X, Z = 2
D, T = 3
L = 4
M, N = 5
R = 6
Insufficient consonants = 0
(e.g., DARLINGTON = D645)

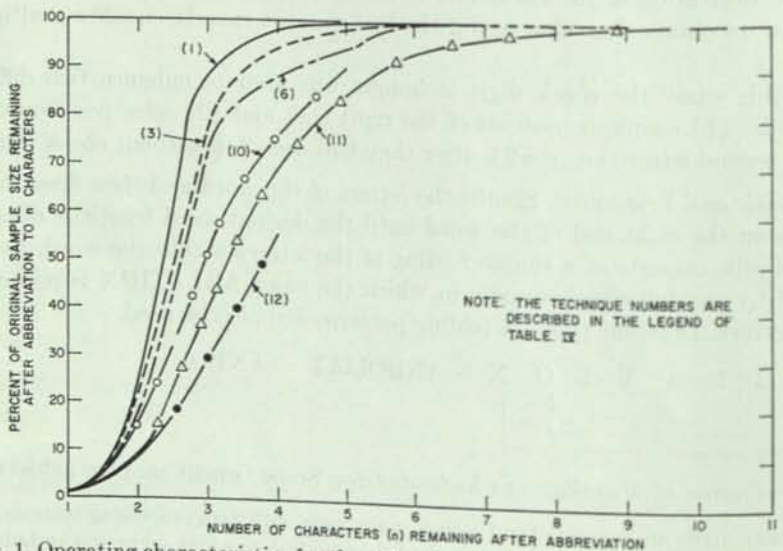


FIG. 1. Operating characteristics for the abbreviation schemes tested with subject words

(4) Special cases:
   (i) If there are insufficient letters, fill out with zeros (e.g., MORAN = M650)
   (ii) Drop out the second letter in a letter pair (e.g., KELLEY = K400)
   (iii) Drop out adjacent equivalent letters (e.g., JACKSON = J250)
   (iv) Drop out adjacent equivalents of the first letter (e.g., LLOYD = L300).

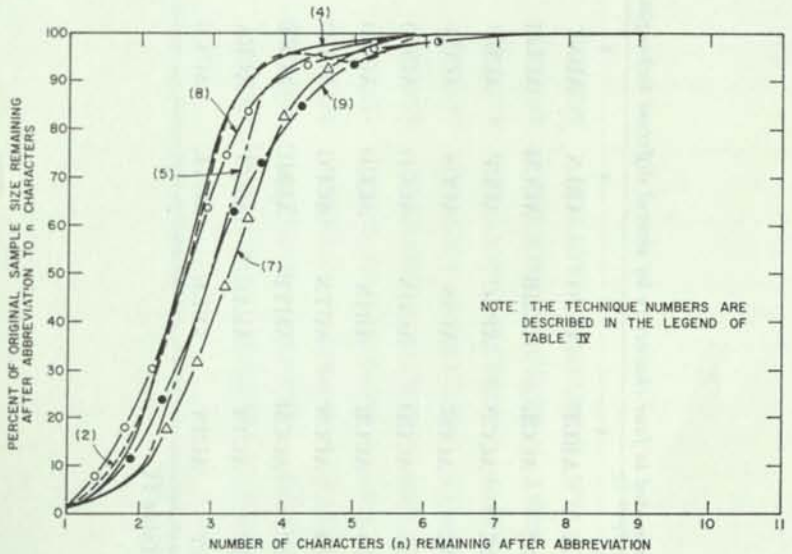


FIG. 2. Operating characteristics for the abbreviation schemes tested with subject words

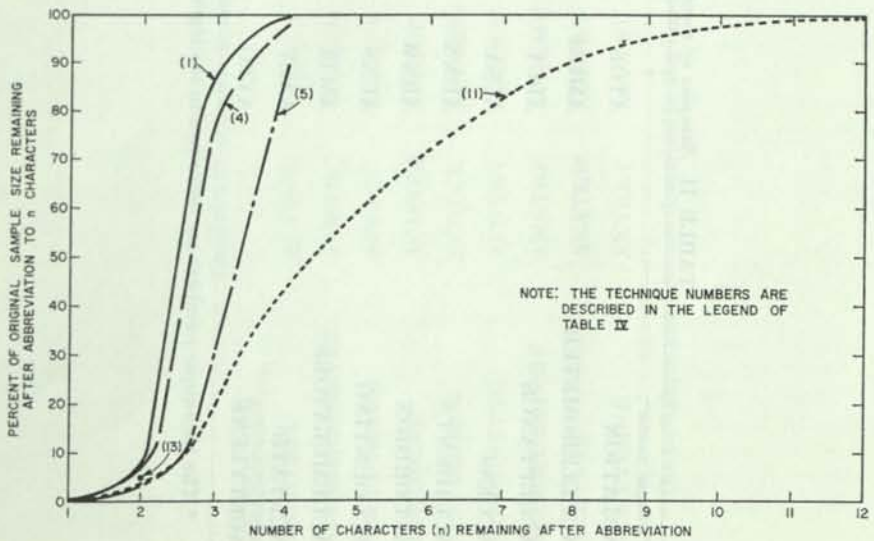


FIG. 3. Operating characteristics for the abbreviation schemes tested with proper names

TABLE II. *Samples of words abbreviated to four characters by several different techniques*

*Technique Number**	1	11	9	4	6	2	8	10
ABLATION	ATOL	ABLA	ABLT	ALTO	ABLN	ABAT	ABLH	ANBO
ACCELEROMETER	AMRM	ACCE	ACCL	ALMR	ACCM	ACLM	ACCM	ARCE
ACCEPTANCE	APCY	ACCE	ACCP	APAC	ACCP	ACPC	ACCM	AECC
ACCESS	ACSA	ACCE	ACCS	ACSS	ACCS	ACCS	ACCR	ASCS
ACCIDENT	ADNN	ACCI	ACCD	ACDN	ACCD	ACCD	ACCA	ATCN
ACCORDION	ARNW	ACCO	ACCR	ARIN	ACCD	ACCD	ACCX	ANCO
ACCOUNTING	AUNS	ACCO	ACCN	AUTN	ACUG	ACUG	ACCW	AGCN
ACCUMULATORS	AMTI	ACCU	ACCM	AMTR	AUMU	ACUU	ACCK	ASCR
ACETATE	AAEV	ACET	ACTT	AEAE	ACTT	ACEA	ACEU	AECT
ACETYLENE	AYEG	ACET	ACTY	AYEE	ACYL	ACYL	ACED	AECN

* The technique numbers are described in the legend of Table IV.

TABLE III. *Samples of words abbreviated to six characters by several different techniques*

*Technique Number**	1	11	9	4	6	2	8	10
ABLATION	ALTONL	ABLATI	ABLATN	ALTION	ABLATN	ABLATI	ABLATM	ANBOLI
ACCELEROMETER	ALMTRG	ACCELE	ACCLRM	ALRMTR	ACCLMT	ACCLRM	ACCELV	ARCECT
ACCEPTANCE	ACPACU	ACCEPT	ACCPTN	ACPACE	ACCPNC	ACCPTC	ACCEPR	AECCCN
ACCIDENT	ACDNTQ	ACCIDE	ACCDNT	ACDENT	ACCDNT	ACCIDN	ACCIDN	ATCNCE
ACCORDION	ACRINK	ACCORD	ACCRDN	ACRION	ACCODN	ACCODO	ACCORQ	ANCOCI
ACCOUNTING	ACUTNV	ACCOUN	ACCNTN	ACUTNG	ACCUNG	ACCOUG	ACCOUM	AGCNCI
ACCUMULATORS	AMLTRE	ACCUMU	ACCMLT	ACMLTR	ACCUMU	ACCUMU	ACCUMC	ASCRCO
ACETATE	AEATEW	ACETAT	ACETAT	AETATE	ACETAT	ACETAT	ACETAZ	AECTEA
ACETYLENE	AEYEET	ACETYL	ACTYLN	AEYENE	ACTYLN	ACEYLN	ACETYK	AECNEE

* The technique numbers are described in the legend of Table IV.

Results

Figures 1 and 2 describe the results of the tests on the sample of subject words. Figure 3 describes the results of the tests on the sample of proper names. Tables II and III show the results of abbreviating subject words to four and six characters, respectively, by each of several techniques.

In general, none of the tested techniques produced abbreviations which could be transformed back into the original words, and none of the techniques produced abbreviations which bore a strong resemblance to the original word. With the exception of the statistical data to describe the nature of the words to be abbreviated, no tables or prior knowledge of the input data was required. That is, no table storage and look-up operations were required.

TABLE IV. *Ranking of techniques by performance on subject words*

	Allowable Field Size Number of Characters			
	3	4	5	6
Technique Number:	1	1	1	1
	3	2	4	2
	4	4	5	4
	6	3	3	6
	2	5	7	3
	8	8	8	5
	9	6	9	7
	10	7	2	9
	5	9	6	8
	11	10	10	11
	7	11	11	
	12	12		

NOTE: The best techniques are at the top of the list.

LEGEND

Technique No.	*Description of Technique*
1	Selective Dropout ($n = 2$) with a check letter
2	Selective Dropout by separate rankings of character usage for each letter position
3	Selective dropout by a single ranking of bigram usage
4	Selective dropout ($n = 2$)
5	Selective dropout ($n = 3$)
6	Selective dropout by a single ranking of letter usage
7	Selective dropout ($n = 4$)
8	Truncate right end and add check letter
9	Vowel elimination
10	Shuffle
11	Truncate right end
12	Truncate left end
13	First letter and last consonant of the edited string of characters (used on names only)

In particular, the familiar techniques of simple truncation and vowel elimination produced the poorest results. The schemes which took advantage of the statistical nature of the words (i.e., drop-out by a composite frequency distribution, drop-out by a separate frequency distribution for each letter position, and drop-out by bigram rankings) produced better results. The shuffle scheme produced mediocre results, and the technique of selectively dropping the *n*th letter produced some very good results and some bad results. Two basic schemes (truncation, selective drop-out of every second letter) were run with and without the generation of a check letter. In both cases, the generation of a check letter improved the performance of the basic scheme.

For subject words and for proper names, the best performance was obtained by dropping every other letter and generating a check letter (i.e., selective drop-out, $n = 2$, with a check letter).

For the subject words, several of the performance curves crossed each other, so that no single "best" scheme was noted. That is, the "best" scheme depended upon how many letters were allowed for the abbreviation. Because of the large number of tests on the subject words, the curves had to be plotted on two different figures. This provided some difficulties in comparing the techniques against each other. For this reason, a separate table (see Table IV) was constructed to show the rankings by performance of all the tested techniques, for abbreviated word sizes of 3, 4, 5, and 6 characters. This table shows that for subject words and an allowable field size of 3 to 6 characters, the best performance was obtained by dropping every other letter and generating a check letter. The test results for proper names did not intersect on the curve, and thus there was a single "best" technique for all degrees of abbreviation. The actual test results for subject words and proper names are shown in Tables V and VI, respectively.

The subject words of the data base ranged in length from 1 to 23 characters.

TABLE V. *Number of unique subject words remaining after abbreviation to n characters*

Number of Characters (n)	*Technique Number**											
	1	2	3	4	5	6	7	8	9	10	11	12
1	26	26	26	26	26	26	26	26	26	26	26	24
2	511	418	472	401	196	388	196	511	300	335	196	183
3	1831	1511	1611	1576	1060	1545	841	1427	1087	1081	841	623
4	2056	1997	1965	1991	1912	1871	1737	1890	1653	1613	1456	1158
5	2078	1960	2043	2054	2048	1957	2011	2006	1968	1886	1762	
6	2080	2077	2069	2075	2068	2073	2068	2046	2051		1938	
7	2082	2081	2076	2077	2077	2079		2072	2078		2012	
8		2082	2082	2079		2081		2078	2082		2054	
9				2080		2082		2082			2073	
10											2081	
11											2082	

* The technique numbers are described in the legend of Table IV.

NOTE: The original list contained 2082 unique subject words.

TABLE VI. *Number of unique names remaining after abbreviation to n characters*

Number of Characters (n)	*Technique Number**				
	1	4	5	11	13
1	25	25	25	25	
2	606	557[a]	245	245[a]	317[a]
3	7117	6313	2559	1542	
4	8122	8013	7115	3561	
5		8171		4914	
6				5875	
7				6756	
8				7377	
9				7766	
10				7953	
11				8042	
12				8113	

* The technique numbers are described in the legend of Table IV.

[a] These schemes are often used for coding names into edge-punched cards.

NOTE: The original list contained 8184 unique names.

However, with most of the abbreviation techniques the data could still be represented by 8 to 10 characters while retaining the same degree of discrimination as the original list. That is, the number of unique file items remained the same after abbreviation as for the original data base, even though the field size had been restricted to 8 or 10 characters. The proper names of the data base ranged from 6 to 22 characters after editing. However, the names could also be reduced to 8 to 10 characters while still retaining a very high percentage of the originally unique file items. These figures will probably vary with the total size and configuration of the data base being processed.

Summary

There is a need for abbreviation techniques which can systematically be applied to a wide variety of source material, and which satisfy many of the general objectives stated earlier. This study described several techniques which can be used for this purpose. In addition, many of the techniques were tested against a large data base in order to provide some quantitative comparisons. The tests seem to indicate that the generation and use of a check letter would enhance almost any of the basic techniques. For subject words and for proper names, the technique of selectively dropping out every second letter, while generating a check letter, seemed to be the most effective. However, these results should not be viewed as absolute findings, and must be tempered by recognition that some described techniques were not tested, and that many other techniques were neither described nor tested. The results and the general approach may serve as the stimulus for additional work on this subject.

REFERENCES

1. Barrett, J. A. and Grems, M. Abbreviating words systematically. *Comm. ACM 3* (1960), 323–324.
2. Bemer, R. W. Do it by the numbers—digital shorthand. *Comm. ACM 3* (1960), 530–536.
3. Blair, C. R. A program for correcting spelling errors. *Inf. Contr. 3* (1960), 60–67.
4. Bloomer, J. G. *Bloomer's Commercial Cryptograph—A Teletype and Double Index—Holocryptic Cipher.* (A. Roman & Co., San Francisco, Calif., 1874).
5. Burton, N. G. and Licklider, J. C. R. Long-range constraints in the statistical structure of printed English. *Am. J. Psychol. 68* (1955), 650–653.
6. Chapanis, A. The reconstruction of abbreviated printed messages. *J. Experimental Psychol. 48* (1954).
7. Cox, G. J., Casey, R. S. and Bailey, C. F. Recent developments in keysort cards. *J. Chem. Educ. 24* (1947), 65–70.
8. Evans, M. W., McElwain, C. K. and Van Hoosen, F. Machine correction of garbled English text. M.I.T. Lincoln Lab. Report, ASTIA Doc. No. AD-237 700 (June 1960).
9. Frishberg, M. Several techniques for obtaining 60% to 150% efficiency improvements in storage and retrieval systems using general purpose computers. Paper presented at the 15th National Conference of the ACM, Milwaukee, August, 1960.
10. Frumkina, R. M. Some procedural problems in compiling frequency dictionaries (statistical structure of dictionary and text). A translation of an article which appeared in the Russian-language periodical *Mashinnye Pereved i Prikladaaya Lingvistika* (Machine Translation and Applied Linguistics), No. 2(9), Moscow (1959). Translation available from Office of Technical Services, U. S. Dept. of Commerce, Document No. JPRS: 3599 (26 August 1960).
11. Gaines, H. F. *Cryptanalysis.* (Dover Publications, Inc., New York, 1956).
12. Griffith, R. T. The Minimotion typewriter keyboard. *J. Franklin Inst. 248*, (1949), 399–436.
13. Korolev, L. N. Coding and code compression, *J. ACM 5*, (1958), 328–330.
14. Luhn, H. P. Superimposed coding with the aid of randomizing squares for use in mechanical information searching systems. In Casey, Perry, Kent and Berry, *Punched Cards—Their Application to Science and Industry*, 2d ed., ch. 23, Reinhold Pub. Corp., New York, 1958.
15. Mandelbrot, B. Simple games of strategy occurring in communication through natural languages. *IRE Trans.* **PGIT-3** (1954), 124–137.
16. Miller, G. A. Some effects of intermittent silence, *Am. J. Psychol. 62* (1957), 311–313.
17. Miller, G. A. and Friedman, E. A. The reconstruction of mutilated English texts. *Inf. Contr. 1*, (1957), 38–55.
18. Miller, G. A., Newman, E. B. and Friedman, E. A. Length-frequency statistics for written English, *Inf. Contr. 1*, (1958), 370–389. (This research was conducted under contract AF (33(038)-14343, and appears as ASTIA Report No. AD-160 709.)
19. Miller, G. A. and Newman, E. B. Tests of a statistical explanation of the rank-frequency relation for words in written English, *Am. J. Psychol. 63* (1958), 209–218.
20. Newcombe, H. B., Kennedy, J. M., Axford, S. J. and James, A. P. Automatic linkage of vital records, *Science 130* (1959), 954–959.
21. Newman, E. B. The Pattern of vowels and consonants in various languages, *Am. J. Psychol. 64*, (1951), 369–379.
22. Newman, E. B. and Gerstman, L. S. A new method for analyzing printed English, *J. Experimental Psychol. 44*, (1952), 114–125.
23. Newman, E. B. and Waugh, N. C. The redundancy of texts in three languages. *Inf. Contr. 3* (1960), 141–153.
24. Oettinger, A. G. The distribution of word lengths in technical Russian. *Mech. Translation 1*, (1954).

25. OETTINGER, A. G. Account identification for automatic data processing. *J. ACM 4*, (1957), 245–253.
26. OHAVER, M. E. *Cryptogram Solving*. (Stoneman Press, Columbus, Ohio, 1933).
27. OHLMAN, H. Subject word letter frequencies with applications to superimposed coding. *Proceedings International Conference on Scientific Information*, Washington, D. C. (November 1958).
28. PRATT, F. *Secret and Urgent, The Story of Codes and Ciphers*. (Blue Ribbon Books, Garden City, N. Y., 1942).
29. REMINGTON-RAND. Soundex—foolproof filing system for finding any name in the file. Brochure LBV809 (undated).
30. REMINGTON-RAND. Idem sonans says it's legal (Soundex). Brochure LBV528 (undated).
31. SHANNON, C. E. Prediction and entropy of printed English. *Bell System Tech. J. 30*, (1951), 50–64.
32. SMITH, L. D. *Cryptography—The Science of Secret Writing*. (W. W. Norton Co., Inc., New York, 1943).
33. TAUNTON, B. W. Name code—a method of filing accounts alphabetically on a computer. *Data Proc. 2*, (March 1960), 23–24.
34. WEST, M. *A General Service List of English Words with Semantic Frequencies*. (Longmans, Green, and Co., New York, 1953).
35. YNGVE, V. H. Gap analysis and syntax, *IRE Trans. Inf. Theory IT-2*, (1956), 106–112.
36. ZIPF, G. K. *The Psychobiology of Language*. (Houghton Mifflin Co., Boston, 1935).
37. ZIPF, G. K. *Human Behavior and the Principle of Least Effort*. (Addison-Wesley Publishing Co., Inc., Cambridge, Mass., 1949).
38. BOURNE, C. P. AND FORD, D. F. A study of the statistics of letters in English words. *Inf. Contr. 4*, 1 (Mar. 1961), 48–67.

DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE

PUBLIC HEALTH SERVICE . BETHESDA 14, MD.

NATIONAL LIBRARY OF MEDICINE

Refer to: NLM-DPD

August 21, 1964

Dr. Charles Bourne
Stanford Research Institute
Palo Alto, California

Dear Dr. Bourne:

The National Library of Medicine in conjunction with the American Standards Association, Z-39 Committee on Library Standards, Subcommittee on Machine Input Records, is interested in investigating any research that may have taken place regarding truncating of personal names for computer purposes.

Specifically, our subcommittee is trying to determine the ambiguity that results after truncating surnames at various levels. For instance, if we were to take all the names in a large directory, perhaps a telephone book, and truncate at various levels, perhaps from 15 down to 5 letters, how many ambiguities would be introduced at each level? If we were to include the first and middle initials, to what degree would the ambiguities be reduced?

Before we begin any research on this particular topic we thought it would be wise to investigate any similar work that has already been done. I would appreciate it very much if you could let me know of any work that you may have done in this general area or any knowledge that you might have of similar work done by others.

Thank you very much for your assistance.

Sincerely yours,

Charles J. Austin

Charles J. Austin
Chief, Data Processing Division
National Library of Medicine

Z-39 Subcommittee on Machine Input Records