Direct coding of English language names

By D. A. Brace

Bookshops send orders to Book Centre Ltd. listing the titles of the books they require but not their catalogue numbers. Before an invoice can be produced by their IBM 1401 computer, the address code of each title must be found manually from an alphabetical list containing 19,000 titles. This paper describes a method by which the computer is programmed to locate the address directly from a punched card carrying an abbreviated form of the title, thus avoiding the manual searching procedure. The interest of the paper lies in the rules for abbreviating book titles to the minimum number of characters consistent with sufficient discrimination between titles and ease of working by normal punch operators.

Book Centre Ltd. on the North Circular Road, London, is a centralized warehouse and despatching organization for 40 publishers. It is operated on a joint basis by the 40 members, and has the form of a co-operative with each publisher taking a share of the annual costs in proportion to his turnover.

Orders for books are received from booksellers and are turned into an invoice by using an IBM RAMAC computing system. However, computers work in numbers and in order to produce an invoice by the RAMAC computer, each title listed on the customer's order must be coded into numerical terms. This coding operation is called *Catalogue Indexing* and is carried out by means of looking up the title in an alphabetical list which gives the *RAMAC disk address*. About 3,000 invoices are produced each day and, since each order averages three items, there are 9,000 titles to be looked up in the alphabetical lists daily. About 20 staff are employed in looking up these numbers.

Tracing numbers through an alphabetical list is a relatively crude operation, and is a serious barrier to an automated system. This is clear when it is appreciated that the computer has a capacity for three times as many publishers as at present and that, if this planned expansion is to be realized, the Indexing operation will require an army of 50–60 staff, a very unsatisfactory situation.

As part of an operational research study being carried out at Book Centre, a method has been proposed by which book titles are punched directly on to a punched card and the computer itself performs the indexing function of locating the reference on the disks. The point here is that since each title is unique a method must exist which will relate the letters comprising the title to the RAMAC address. The system has not yet been programmed and installed, and this article describes the first part of the project, in which the reduction of English words to a basis suitable as a computer code was developed.

Abbreviating the title

If a substantial economy in staff is to be achieved, the time taken by the punch girl in entering the title code directly on to a card must not be very much greater than her time to enter a RAMAC Index number, and so the title must be abbreviated to the minimum number of letters which will still identify it accurately.

The problem reduces therefore to finding a systematic method of abbreviating titles which balances the need for extreme brevity with the difficulty of maintaining uniqueness, or at least distinctness. Whatever method is proposed is further subject to the constraint that the punch girl should be able to reduce the title to its abbreviated form without serious diminuition in punching speed, using a document that is normally in handwriting and often very poorly presented.

No system of abbreviation can ever guarantee perfect discrimination for every possible title, and it is inevitable that some clashes will occur; these are treated by special methods.

Indirect addressing

In order to explain the automatic indexing procedure, it is best to describe first the technique of *Indirect Addressing*. This is a sophisticated but well-known computer technique and comprises a substantial part of the indexing procedure.

Suppose that the book trade ordered by catalogue number rather than by title. The order would then be already encoded, and each item ordered would be described by a unique number. The number would not be a RAMAC address, but a program could be devised which would convert this code to an address. The punch girls would put the number directly on to a card without any indexing process.

This conversion program is quite common in computing —for instance, where engineering spares are described by a part number—and is called *Indirect Addressing*. A detailed description may be found, for instance on pages 50 to 62 of the RAMAC 305 Manual issued by IBM. Suppose the catalogue number is of the form AM/5904/B61. This collection of letters and digits must be converted to a RAMAC address which is a five-digit number between 00000 and 50,000. The usual process is to "shuffle" the characters so that they are thoroughly randomized, subtract a constant from them and then divide by another constant so the codes are spread evenly over 0–50,000. Inevitably, this process causes some confusion of addresses, and catalogue numbers that were initially distinct become the same address number. However, the pairs of codes that are confused by this process are always the same, and so the machine compares the

catalogue number on the card with the number in the address; where this differs, the machine passes to a further address linked to the first address, until the catalogue number on the card is found.

Coding of titles

The previous paragraph described how the techniques for converting any catalogue number, whether in letters or numbers, are well established in computer programming. However, the book trade is not in the position that booksellers themselves find catalogue numbers, and the problem therefore remains of converting the title in words to a brief code that the computer can handle, which is distinct for every title.

The problem of encoding a title direct may be considered in three stages. There is the title as it is normally understood—*Wind in the Willows*, *Wuthering Heights*; it is this "stem" of the title that is to be encoded. Secondly, there is what may be called the "sub-title", such as "Volume I, School Edition, Paperback, With Answers, Notes, Vocabulary, 1890–1914, Cheap Edition, Cloth Boards, Limp, Book XIV". The subsidiary descriptions are vital in defining the book precisely within any one title, and are more varied in form than are "title stems" which are usually one to four common words. At the worst, the volume may be distinguished by its price, particularly in very popular books such as *Toad of Toad Hall* which may run to a dozen different classes. Thirdly, the actual order as received from the bookseller may differ to a lesser or greater extent from the title as printed on the book. This difference may be regarded as a "band of error" about the real title, and ideally the coding system should cope with the more frequent types of error. The most common single error is omission of the name of the author.

Title stem

The coding of titles reduces to deciding how many letters to select from a title and which letters these shall be. The present codes are of five digits, and since these are perfectly efficient, i.e. each number is used exactly once, any coding system must use more than five digits. However, five letters may be quite adequate since letters contain more information than numbers as they distinguish 26 cases instead of ten.

If one letter is used—say the first letter of the title—the number of titles that can be distinguished is at most 26. In practice the number is less than this since some letters are far more common than others. If two letters are used the number of possible cases is $26 \times 26 = 676$. Three letters give about 18,000 possible cases, and four letters half a million.

The total number of titles in existence is a quarter million, and in Book Centre about 19,000. However, booksellers are accustomed to stating the publisher on their orders, and this extra information permits the selection to be made only within the publisher's list. Book Centre publishers have at most 3,000–4,000 titles, although larger publishers may become members in the future.

With so small a number of titles, three code letters should be sufficient since these provide up to 18,000 possible codes. However, there are many letter combinations and letters that rarely occur, e.g. such as z, q, j, k, x, y; also we find that book titles tend to concentrate on a few specialized words so that it is difficult to distinguish between titles with a very brief code. We conclude below that a five-letter code is required.

In choosing which letters to select, an exploration of the theory of linguistics and information theory has discovered only one theorem which seems to be of value; namely, that since in English the second letter of a word tends to be one of the five vowels, little information is obtained from this letter. Other letter positions appear to provide equal amounts of information, and any of them may be used.

We are severely constrained in choosing letter positions by the requirement that punch girls are to select the letters visually—i.e. it would be useless to require a punch girl to select the seventh letter of a word. The obvious letters are the first and third, particularly since these are available for all words above two letters. We therefore select for further study code systems using only these two letter positions.

Trials with Methuen list

The Methuen list ot titles was selected for experiment since at about 2,800 titles it is larger than any other Book Centre publisher except Pitman, so that any coding system adequate for this list will certainly be adequate for any smaller publisher, whose titles will be spread less densely over the alphabet. Pitman, with a larger list,

Table 1

Methuen list: initial letters of titles (approximate) Descending Order

	INITIAL LETTER	NO. OF TITLES		INITIAL LETTER	NO. OF TITLES
1	S	260	14	F	110
2	T	230	15	R	100
3	C	210	16	D	95
4	P	210	17	W	90
5	L	190	18	N	70
6	E	160	19	U	40
7	M	160	20	J	35
8	H	150	21	K	30
9	I	130	22	V	30
10	A	130	23	Y	15
11	B	130	24	Q	10
12	G	120	25	X	5
13	O	120	26	Z	2
					2832

was excluded, since a system adequate for the largest publisher of all would provide extravagantly good discrimination for the smaller publishers. If much larger publishers become Book Centre members in the future, additional treatment may prove to be necessary.

The number of titles beginning with the various letters of the alphabet are given in Table 1; there is a great variation in the frequency of the different letters, and similar variation occurs in the other letter positions.

We selected for detailed study books beginning with A and S—A being an average letter at tenth position and S being the most common and hence providing the greatest probability of clashes; this probability rises rapidly as the "density" of the letter increases.

The number of words in titles is approximately as shown in Table 2; this excludes all words of one or two letters and "and, the, for, from" and not counting "Subtitles". This table shows that two or three words is the general form for a title.

This distribution suggests that a code of at least two letters will be required for one-word titles (676 maximum number of cases), of three letters for two-word titles (18,000 cases), and of four letters for the remainder. Such a code would be as follows:

1-word titles	letters 1 and 3;
2-word titles	first word—letters 1 and 3 second word—letter 1
3-word or more titles	first word—letters 1 and 3 second word—letter 1 third word—letter 1

At this stage the problem of sub-titles is disregarded and authors' names are not used.

Some titles from the S-group are reproduced in Table 3, and the method can be followed through using the following as examples:

TITLE	CODE
SACRAMENTS	SC
SACRED WOOD	SCW
SAGA OF THE SERGEANT	SGS
SAILING ALONE ROUND THE WORLD	SIAR

Table 2

NO. OF WORDS	PERCENTAGE OF TITLES	IN 2,800 TITLES	
1	7	190	
2	46	1,300	
3	34	960	1,310
4	9	260	
5	2	60	
6	1	30	
		2,800	

A number of common words are excluded, namely, *all one and two-letter words*, and also "and, the, for, from".

The result of this code system is not satisfactory. In the 130 titles beginning with A there were nine clashes, i.e.

NO. OF WORDS (EXCLUDING SUB-TITLES)	NO. OF TITLES	NO. OF CLASHES
1	9	3
2	54	6
3+	67	0

The author initial letter was therefore added to the one and two-word titles to bring them to a four-letter code, thus:

TITLE AND AUTHOR	CODE
SACRAMENTS—Oxenham	SC—O
SACRED WOOD—Elliot	SCWE
SAGA OF THE SERGEANT—Ivanov	SGSI
SAILING ALONE ROUND THE WORLD—Slocum	SIAR

With this code all titles of two or more words have four letters and must be judged together. The results for the 260 titles beginning with S are:

NO. OF WORDS	NO. OF TITLES	NO. OF CLASHES	
1	17	0	5% clashes
2+	245	12	

The results with this code and with minor variations of it are 5–7½% clashes with the S-group—the most difficult block—and 2–5% for the A-group, which is the typical block; this improves on the previous code, but is still inadequate. The reason for this relatively poor result is not hard to find. Publishers have a tendency to choose title words from an extremely restricted set, possibly because they have a familiar ring. Out of the 262 titles beginning with S, no less than 20 begin with "Selections-", a dozen with "Shakespeare" and substantial blocks with "Second, Secondary", "Story of-" and "Secret". Other letters show the same phenomenon, notably books called "History of————", etc.

The titles:

Selections from the Brownings
Selections from Bunyan
Selections from Byron

all code as SLB— under this rule (the dash signifying no author).

A high degree of correlation between title-words therefore clearly exists, and a five-letter code cannot be avoided. The final code would then be four letters from the title, and a fifth letter either signifying the author or specifying the sub-titles as described in Table 4. In addition, the publisher must be specified and this might be by using a single letter code, since although there are at present 40 publisher-members, some are quite small and can be grouped together under the same letter.

Empty columns are left in the codes for one-word titles

Table 3

Methuen Titles : Beginning of S section

46883	SACRAMENTS	OXENHAM	3d.
46884	SACRED WOOD	ELIOT	10/6
45476	SACRED WOOD Paper	ELIOT	7/6N
46885	SAGA OF THE SERGEANT	IVANOV	4/6
45427	SAILING ALONE ROUND THE WORLD Vent Lib.	SLOCUM	5/-
46886	SAILING IN A NUTSHELL	BOYLE	8/6N
47462	ST GEORGE & THE DRAGON	JOHN	12/6N
46887	SAINT JAMES IN SPAIN	KENDRICK	25/-N
46889	ST. MARK	RAWLINSON	25/-N
46891	SALTHAVEN	JACOBS	6/-N
46892	SAM THE SUDDEN	WODEHOUSE	6/-N
47463	SAMSON'S BREAKFAST	MAKOWER	8/6N
46893	SAVAGE AFFAIR	SCOTT	5/-
46895	SCANDAL IN TROY	HANSEN	5/-N
46896	SCARLET U.	MATTHIESSEN	10/6N
46370	SCHACHNOVELLE 24/5/62	ZWEIG	6/-
46899	SCHOOLS OF PAINTING 10th Edn. RP. 1953	INNES	18/-N
46900	SCHOOL SERVICE Clth Boards	WAYNE	7/6
46901	SCIENTIFIC STUDY OF SOCIAL BEHAVIOUR	ARGYLE	25/-N
45335	SCULPTURE THROUGH THE AGES Outline	HAGGAR	10/6N
46903	SEA BATTLE	BIRMINGHAM	6/-N
46904	SEA LADY 9th Edn. 1951	WELLS	7/-N
46906	SEAFARER, THE	GORDON	7/6
46907	SEALSKINS FOR SILK	CHEESMAN	12/6N
47464	SEASON OF MISTS	TRACY	15/-N
46910	SEA WHISPERS	JACOBS	7/-N
45055	SECOND BOOK OF NAUGHTY CHILDREN	BLYTON	8/6N
45082	SECOND FORM AT MALORY TOWERS	BLYTON	9/6N
45063	SECOND FORM AT ST. CLARES	BLYTON	9/6N
46911	SECOND YEAR GERMAN	STOCKTON	4/-
47465	SECONDARY CERTIFICATE QUESTIONS English Language		4/6
46912	SECONDARY SCHOOLS SELECTION	VERNON	15/-N
46913	SECRET AGENT	CONRAD	15/-N
46914	SECRET BATTLE	HERBERT	6/-N
47468	SECRET LANGUAGE	NORDSTROM	10/6N
46915	SECRET POWER	CORELLI	9/6N
46916	SECRET OF RUSTCOKER	GREEN	3/6N
45375	SECRET OF THE UNICORN		8/6N
46890	SEE HOW THEY WORK. BK. 2 5/7/62	BERG/CLARK	10/6N
45229	SEISMOLOGY	BULLEN	10/6N
46917	SELECTED ESSAYS OF HILAIRE BELLOC		12/6N
46918	SELECTED ESSAYS OF G. K. CHESTERTON		18/-N
46919	SELECTED ESSAYS OF E. V. LUCAS	WETHERED	10/6N
46920	SELECTED POEMS OF VICTOR HUGO Sch. Edn.		8/6N
46933	SELECTIONS FOR SECONDARY EDUCATION		9/6N

Table 4

TITLE and AUTHOR	CODE
SACRAMENTS—Oxenham	SC——O
SACRED WOOD—Elliot	SCWOE
SAGA OF THE SERGEANT—Ivanov	SGSRI
SAILING ALONE ROUND THE WORLD—Slocum	SIARS

(or another spacing symbol) which increases the discrimination between titles of different lengths.

This final code produces only two clashes in the S-group and, since this is the largest group, the total clashes in Methuen as a whole are small. The clashes are:

Selections from Borrow (no author)		=SLBR—
Selections from Byron (no author)		=SLBR—
Short History of the Roman Empire to the Death of Marcus Aurelius	—Wells	=SOHRW
Short History of Rome to the Death of Augustus	—Wells	=SOHRW

Discrimination in such extreme cases as these is virtually beyond the reach of any brief coding system. Special methods are used to cope with these clashes.

Sub-titles

The coding rules should, if possible, utilize the author's name and any sub-title information supplied, since this is all discriminating information; but booksellers often omit the author's name, and sub-titles occur only with a minority of titles, and yet the coding should not break down when these are absent. This difficulty can be overcome by treating the five-letter code in two classes.* The first four letters should be the normal code derived from the title stem only. This code will be sufficient in the majority of cases to identify the book completely, and will be passed through the indirect-addressing procedure to find the RAMAC address. In the remaining minority of cases, the machine will note that the code still relates to several titles, (for example, the different editions of the same book). In these cases the fifth letter will be inspected to decide the book.The fifth letter will normally be the author letter. The machine will only fail to find the precise book when both the bookseller has omitted the author and the book is not unique under the four-letter code; the coincidence of these cases will be rare. Thus dividing the code letters into two classes provides the flexibility in the coding necessary to cope with variation in titles and in booksellers' methods.

Thus the reference is divided into two classes—required and supplementary information. This permits reaching the correct address immediately, even though the author's name or sub-title is omitted whenever this supplementary information is redundant, which it is in the majority of cases. The process can be likened to looking up a name in a telephone directory when the surname is known but the initial is in doubt. If only one person of that name is present in the directory the number is found immediately, while if the initial had been wrongly included as an integral part of the search argument, the number would not have been found. If there are only two or three names, the hazy knowledge of the initial will often be adequate.

The fifth letter in the code will be the author letter when this exists, unless there is other sub-title information, in which case this latter takes precedence over the

* I am indebted for this suggestion to Mr. C. Vince of the Programming Support Group of IBM.

author. Preliminary study suggests that there are not more than a score of sub-titles of importance, e.g. "Volume I", "paperback", "school edition", etc., and a regular code is being specified which will cover the bulk of these cases.

Duplication of codes

It is anticipated that some errors will occur because of duplication of codes and for other reasons, and it is therefore not intended at present that invoices will be produced directly from the coded title cards, but that the list of titles will be printed out first for visual verification against the original order. The final invoices are already checked against customers' orders in the present procedure, so that no extra costs will be caused by the new process—the checking operation is merely brought forward to an earlier stage.

As well as the printed list, a new input punched card will be produced, with the machine address number on it in place of the encoded title. Any cards found to be in error after checking the printed list can then be corrected.

Alternatively, either the input data or the output data may be written electronically directly onto the disk storage. Then, after correction cards have been inserted to overwrite the listings that are in error, the invoices may be produced automatically without further input.

In the cases where there is a duplication of codes, the two or more titles will be listed on the print-out but the computer will record only the more frequently demanded of the two clashing titles every time. Therefore the use of a correction card will be necessary only when the lesser-used title is required. For instance it turns out that neither of the two titles *Selections from Borrow* and *Selections from Byron* has been required very often, so that the fact that their codes are duplicated would not cause much difficulty in practice.

This illustrates the point that the frequency of clashes over the whole list of titles is only a theoretical measure of the effectiveness of the coding system, and assumes implicitly that each title is required with equal probability. This is the least favourable assumption for the coding system, and in practice the variation in the probability with which different titles are required should be taken into account. On average, by so arranging that the computer always chooses the more frequently demanded of two clashing titles, the number of times that corrections will have to be made will be much less than that suggested by the proportion of titles that are confused.

Punching speed

The system suggested here has not yet been installed, and the reduction in punching speed and the loss of accuracy that may be caused by requiring punch girls to combine the relatively intellectual occupation of deriving a title code with the mechanical process of card punching has not yet been assessed. Experiments will be made to measure these factors, and possibly a redundant coding, such as punching the first three letters of every word, might prove to be quicker to use. However, because the normal procedure of verifying the punched card will be discontinued, since a quicker, visual verification occurs at the machine stage, a moderate reduction in punching speed will not increase the cost of the operation.

Acknowledgements

I am indebted to Mr. F. D. Sanders of Book Centre Ltd. for permission to publish this paper. On the matters of information theory, I have been greatly helped by Mr. R. A. Fairthorne of R.A.E., Farnborough, who has been extremely generous with his time.

Reference

BOURNE, Charles P., and WARD, Donald F. (1961). "A study of methods for systematically abbreviating English words and names", *J. Assoc. for Computing Machinery*, Vol. 8, No. 4. (This article contains a bibliography of 38 references.)

Book Review

"Microelectronics Using Electron-Beam-Activated Machining Techniques" by K. R. Shoulders. (Part of *Advances in Computers*, Vol. 2, Academic Press, 1961.)

The author starts from the assumption that a computer of the future is likely to need 10^{11} or more electronically active components, and his purpose is to outline a programme of research work which may ultimately lead to the fulfilment of this requirement. He concludes, as others have done, that the most promising method of fabricating the components is by vacuum evaporation, and his programme proposes the use of active elements depending on the field-emission of electrons. An eventual packing density of 10^{11} components per cubic inch is postulated, and problems of heat dissipation and the inter-connection of components are discussed. A large part of the article is devoted to techniques of fabrication, including methods of micro-machining, the production of etching resists, the preparation of substrates, the control of evaporation and the examination of the final product by electron-optical methods. Finally, there is a description of ultrahigh-vacuum apparatus which is being built for these purposes.

This is a stimulating article which anyone interested in these matters can read with profit. It is also an infuriating article which continually leaves the reader in doubt whether a particular statement refers to something which has already been achieved, to a hope not yet fully realized, or to a pipe dream which is unlikely to come to pass within the next twenty years.

C. W. OATLEY