Patent Office
Research and Development
Reports . . . . . No. 15

A NOTATION SYSTEM FOR TRANSLITERATING TECHNICAL AND SCIENTIFIC TEXTS FOR USE IN DATA PROCESSING SYSTEMS

Prepared by

Simon M. Newman and Rowena W. Swanson
Office of Research and Development
Patent Office

Kenneth Knowlton
Research Laboratory of Electronics
Massachusetts Institute of Technology

Office of Research and Development

Patent Office

May 22, 1959

U. S. DEPARTMENT OF COMMERCE
Lewis L. Strauss, *Secretary*

PATENT OFFICE
Robert C. Watson, *Commissioner*

For sale by the U. S. Department of Commerce, Washington 25, D. C. Price 25 cents.

Preface

This notation system provides a set of characters by which texual material, particularly the contents of scientific and technical texts, can be represented. The system has been initially designed to accommodate the complexities of the texts of United States patents. Symbology has been devised to preserve details which we believe will eventually be needed from patent texts. Conversely, techniques have been incorporated in the notation system for producing uniformity in certain sectors of patent text where such uniformity appeared preferable. These diverse demands of detail and uniformity have apparently not been required in the projects described in items of the bibliography.

While it is planned to use the transliterated patent corpus for studies of retrieval of information from raw text, the corpus will first be used by Kenneth Knowlton to create a record tape of a format which can be utilized in linguistic analysis programs. Since similar work has been done by MIT with other corpora, existing programs are serving as springboards for the creation of newer programs. These are described in Section 5.

The notation system is amenable to alteration to suit purposes other than that of its first use, including transliterations of non-patent, scientific and technical texts. The system will undoubtedly be changed and enlarged, foreseeable additions including procedures and nomenclature for representing the elements of mathematical and chemical formulas.

The undersigned is indebted to his co-authors--to Ken Knowlton who established the basics of the system and to Rowena Swanson who weeded, amplified, and edited. Thanks are also due to Dr. Victor H. Yngve of MIT and to Don D. Andrews, Director of the Office of Research and Development of the U. S. Patent Office, for helpful suggestions.

Simon M. Newman

Table of Contents

LIST OF TABLES AND FIGURES

A NOTATION SYSTEM FOR TRANSLITERATING TECHNICAL AND SCIENTIFIC TEXTS FOR USE IN DATA PROCESSING SYSTEMS

1.0 INTRODUCTION

This report provides--

a. the codes and instructions for punching (and verifying) certain portions of the headings and all of the specifications, claims, and cited references which are contained in the texts of U. S. patents, and

b. a comparison of the punched card format with the IBM 704 tape output format.

The notation system which has been devised for preparing the punched card text is *machine oriented* to simplify the program for creating a magnetic tape from the card deck. This imposes a greater load on the card puncher since mnemonic codes and other memory aids could but infrequently be incorporated. The IBM 704 tape output is *user oriented* and is designed to be readable by many who are not familiar with the notation system.

The codes in Sections 2 and 3 are expressed in IBM 024 alpha-numeric characters since the IBM 024 Card Punch is being used in the punching operation. Since the Fortran Punch is normally used in conjunction with the IBM 704 in which the punched cards will be processed, the characters of both the IBM 024 and Fortran Punches are listed in the Symbol Dictionary of Section 4, and the Fortran characters, exclusively, are referred to in Section 5 relating to the IBM 704 programs.

The punched text, the magnetic tape, and the programs for creating the tape and for using it to locate and list individual words and groups of words in their context will be used for research in language and in patent searching at our Office of Research and Development, at the National Bureau of Standards, and at the Massachusetts Institute of Technology.

Every effort will be made to make available copies of the card deck, the IBM 704 program decks, and the IBM 704 tape at cost to bona fide researchers.

2.0 SYMBOLOGY FOR TECHNICAL TEXT TYPOGRAPHY

2.1 The codes are expressed in terms of the characters of the standard IBM 024 alpha-numeric Card Punch. The IBM 024 characters are the same as those of the Fortran Punch which is normally used in conjunction with the IBM 704, with the exception of five characters. The punch patterns for these characters and their designations in IBM 024 and Fortran symbology are given in Table 1.

Table 1.--Character Equivalents

PUNCH	ST'D IBM 024 SYMBOL	IBM 704-FORTRAN SYMBOL
8-3	#	=
8-4	@	-
12	&	+
12-8-4	¤	)
0-8-4	%	(

2.2 The card format is as follows:

Columns 1-72 are for the patent text.
Column 73 is left blank.

Columns 74-80 will be used *later* for consecutively numbering the cards. These columns are allocated as shown in Table 2. Column 80 is left blank so that a card, later made and inserted between two existing cards, can be given a decimal designation.

			Col.	74	75	76	77	78	79	80
Thus card	18	is punched						1	8	
card	3,274	is punched				3	2	7	4	
and card	54,875	is punched			5	4	8	7	5	

Table 2.--Card Numbering Format

Col.	74	75	76	77	78	79	80
	100,000's	10,000's	1,000's	100's	10's	units	blank

A new card is begun for each document. Column 1 of each following card contains whatever text consecutively follows column 72 of the immediately preceding card; e.g.:

(1) A word which does not end in column 72 continues in column 1 of the next card:

```
1                                   72
|                                   |
.........................THE OPERATI

ON CALLED PUNCH.........................
```

(2) If a word or sentence ends in column 72, a space is left in column 1 of the next card:

```
1                                   72
|                                   |
............................THE OPERATION

 CALLED PUNCH...........................
```

(3) An abbreviation which does not end in column 72 continues in column 1 of the next card:

```
1                                   72
|                                   |
.........................THE END IS ETC*

*.. *BEGIN NEW SENTENCE................
```

2.3 Letters

Letters in lower case are punched normally, e.g.:

small is punched SMALL

Letters in upper case are introduced by an asterisk (*), a * being punched in front of *each* upper case letter, e.g.:

Reissue is punched *REISSUE

point A is punched POINT *A

Exception 1: A word or words in the text which are entirely capitalized are introduced by the ***begin capitalized words*** symbol

**%

The letters of the capitalized word or words are punched as ***lower case*** letters and the word or words are followed by the ***end capitalized words*** symbol

**¤

Thus, the text

The NEEDLE VALVE is emphasized.

is punched

*THE **%NEEDLE VALVE**¤ IS EMPHASIZED.

Exception 2: The complete title of the document and subtitles in the text are punched in ***lower case*** letters regardless of how they appear in the document. [See Sections 3.1; 3.4].

2.4 All Arabic numbers are punched normally, regardless of the type in which they appear in the text.

Exception: When a phrase or sentence is in a form of type other than normal type, numbers which are included in the phrase or sentence are in the type in which they appear in the phrase or sentence.

Thus, numbers in bold face in normal text are punched normally, e.g.:

Apertures between needle valve **101** and port **102**

is punched

*APERTURES BETWEEN NEEDLE VALVE 101 AND PORT 102

Numbers in italics within a phrase in italics are punched as being in italics, e.g.:

In example 2, the number is in the phrase.

is punched

*#*IN EXAMPLE 2*$, THE NUMBER IS IN THE PHRASE.

[See Section 2.9].

2.5 When a Roman numeral occurs, it is introduced by the ***begin Roman numeral*** symbol

*/

the equivalent Arabic numeral is punched, and the numeral is followed by the ***end Roman numeral*** symbol

*,

as follows:

the post XII is

is punched

THE POST */12*, IS

2.6 A space in the text appears as a space (no punch) in the punched card. With few exceptions, ***NO space appears in the punched card except in correspondence with a space in the text.*** Thus, the text

> Word is separated from word by space, and numbers 1/2, 21, and 4,185 are separated as shown. Sentence is separated from sentence by single space.

is punched

```
1                                                                      72
|                                                                      |
*WORD IS SEPARATED FROM WORD BY SPACE, AND NUMBERS 1/2, 21, AND 4,185 AR

E SEPARATED AS SHOWN. *SENTENCE IS SEPARATED FROM SENTENCE BY SINGLE SPA

CE.
```

Note that a punctuation mark may be part of a word as well as a device for separating words. In the above illustration, a space follows the commas separating the complete numbers 1/2 and 21, but there is no space in 4,185 because the comma is part of the single whole number.

Also note that a *single* space may be "punched" after the period ending a sentence, whereas two spaces frequently appear in printed text.

A "space" consists of any number of spaces from 1 through 71, as long as no other character intervenes. However, a "space" consisting of 72 or more consecutive spaces signifies the end of a document. [See Section 3.8 .

2.7 The period performs a variety of functions in text. These functions are separately noted in the punched text as follows:

2.7.1 The period at the end of a sentence is punched immediately after the last letter of the last word without a space and it is followed by a space, e.g.:

The machine stops. Begin next sentence.

is punched

```
*THE MACHINE STOPS. *BEGIN NEXT SENTENCE.
```

2.7.2 The period as a decimal point is punched between the two appropriate digits of the number with no space before or after it, e.g.:

103.50 is punched `103.50`

2.7.3 The period at the end of an abbreviation is represented by two asterisks punched between the last letter of the abbreviation and the period without spaces, e.g.:

etc. is punched `ETC**.`

2.7.4 For the period at the end of an abbreviation which is the last word of a sentence, an indication of both functions of the single period is necessary. Thus, although only one period occurs in the text, a period for the end of the abbreviation and a period for the end of the sentence are both punched without any intermediate space, e.g.:

The last word is etc.

is punched

```
*THE LAST WORD IS ETC**..
```

2.7.5 The period after a number which begins a paragraph is treated as the period of an abbreviation; e.g., the following text appearing in column 11 of patent 2,709,339

at line 19 We claim:

at line 20 1. In a two stage

at line 51 2. In a two stage

is punched

```
*11@19 *WE CLAIM**C

*11@20 1**. *IN A TWO STAGE

*11@51 2**. *IN A TWO STAGE
```

[See Section 3.5 for patent paragraph and patent claim notation symbology].

2.7.6 Three successive periods (or asterisks) which signify an ellipsis (an omission of words from the text) are represented by the *ellipsis* symbol

```
**H
```

When the ellipsis occurs at the end of a sentence, the ellipsis and the end of the sentence are both indicated, as follows:

```
**H.
```

Thus the text

a, b, c,***n equals u, v, w,....

is punched

```
A, B, C,**HN EQUALS U, V, W,**H.
```

2.8 Bold-face type is introduced by the *begin bold-face type* symbol

```
**#
```

and is followed by the *end bold-face type* symbol

```
**$
```

In the following illustration, the word BOLD appeared in bold face in the text

```
*THE **#BOLD**$ LETTERS STAND OUT.
```

[See Section 2.4 concerning numerals in bold face].

2.9 Letters or words in italics or letters or words which are underscored are preceded by the *begin italics* symbol

```
*#
```

and are followed by the *end italics* symbol

```
*$
```

One letter, one or more words, a whole sentence, or a whole paragraph may be in italics or may be underscored. The ***begin italics*** and ***end italics*** symbols are punched at the beginning and end, respectively, of each consecutive grouping of italicized or underscored letters (the letter, word, sentence, or paragraph). Insertion of a word not in italics or not underscored ends the italicized or underscored grouping.
[See Section 2.4 concerning numerals in italics].

Ex. 1—The phrase "the need is great" is punched

`THE *#NEED*$ IS GREAT`	When only the word ***need*** is in italics
`THE *#NEED*$ IS *#GREAT*$`	When both the words ***need*** and ***great*** are in italics
`*#THE NEED IS GREAT*$`	When the ***entire phrase*** is in italics

Ex. 2—<u>sn</u>ow, not sow

is punched

```
S*#N*$OW, NOT SOW
```

2.10 Quotation marks are replaced by the ***begin quotation*** symbol **Q and the ***close quotation*** symbol **U at the beginning and end of a quotation, respectively, e.g.:

"This sentence is in quotes."

is punched

```
**Q*THIS SENTENCE IS IN QUOTES**U.
```

When a close quotation mark occurs at the end of a sentence, printers usually put the punctuation inside the quotation mark, thus:

"...to the end."

In punching, it is necessary to punch the close quotation mark before the punctuation mark, as though the text read

"...to the end".

so that the punctuation mark will occur at the end of the sentence.

2.11 The same symbol is used for the prime mark, the apostrophe, and the single quote mark. The ***prime-and-apostrophe*** symbol is

**A

Double and triple primes are represented by the appropriate repetition of the prime-and-apostrophe symbol, as follows:

single prime	**A
double prime	**A**A
triple prime	**A**A**A

The prime-and-apostrophe symbol is applied to single quotation marks which both begin and close the quoted text.
The following examples are illustrative.

Ex. 1--the metal's expansion characteristics

is punched

```
THE METAL**AS EXPANSION CHARACTERISTICS
```

Ex. 2--the area of port 94' is

is punched

```
THE AREA OF PORT 94**A IS
```

Ex. 3--the square a'B"c"'D

is punched

```
THE SQUARE  A**A*B**A**AC**A**A**A*D
```

Ex. 4--the "pump handle 'A' "

is punched

```
THE **QPUMP HANDLE **A*A**A**U
```

2.12 Letters or numbers frequently appear in technical texts in superscript or in subscript. The superscript is introduced by the ***superscript*** symbol *& as follows:

arm^{8}, $lever^{B}$, $10^{5.3}$, or $shaft^{aB}$

is punched

```
ARM*&8, LEVER*&*B, 10*&5*&.*&3, OR SHAFT*&A*&*B
```

The ***superscript*** symbol followed by the numeral ***zero*** is used to designate the degree or temperature mark or the angle mark, e.g.:

water at 23° C is punched `WATER AT 23*&0*C`

The subscript is introduced by the ***subscript*** symbol *@ as follows:

cam_{A}, H_2O; H_f, Fe_{2-3}, or $line_{Ba}$

is punched

```
CAM*@*A, *H*@2*O, *H*@F, *FE*@2*@@*@3, OR LINE*@*B*@A
```

2.13 Letters of the Greek alphabet are introduced by the following symbols:

lower case Greek letter symbol **Y

upper case Greek letter symbol **Z

The appropriate Greek letter symbol is punched preceding ***each*** Greek letter which occurs in the text. Normal Roman letters in lower case type are substituted for the Greek letters in punching. (This substitution is called transliteration). The Greek letters and their Roman letter equivalents are given in Table 3.

Exemplary contexts including Greek letters are punched as follows:

$2\pi r$	is punched	2**YPR
$2\Delta\pi r$	is punched	2**ZD**YPR
α numeric	is punched	**YA NUMERIC
$\Sigma\theta^2\sigma_\rho$	is punched	**ZS**YJ*&2**YS*@**YR

Table 3.--Greek Letters and Equivalents

Word Equivalent	Greek Letters		Roman Letter Equivalent
	Upper Case	Lower Case	
alpha	A	α	A
beta	B	β	B
gamma	Γ	γ	G
delta	Δ	δ	D
epsilon	E	ϵ	E
zeta	Z	ζ	Z
eta	H	η	H
theta	Θ	θ	J
iota	I	ι	I
kappa	K	κ	K
lambda	Λ	λ	L
mu	M	μ	M
nu	N	ν	N
xi	Ξ	ξ	X
omicron	O	o	O
pi	Π	π	P
rho	P	ρ	R
sigma	Σ	σ	S
tau	T	τ	T
upsilon	Υ	υ	U
phi	Φ	ϕ,φ	F
chi	X	χ	C
psi	Ψ	ψ	Y
omega	Ω	ω	Q

2.14 For ready reference, IBM 024 and Fortran notations for punctuation marks and other symbols, including the notations discussed in the foregoing subsections and notations not heretofore presented, are summarized in Table 4.

Table 4.--Notations for Typographic Marks

MARK	ST'D IBM 024 SYMBOL	IBM 704-FORTRAN SYMBOL
. (period)	See Section 2.7	
, (comma)	,	,
((begin parentheses)	%	(
) (close parentheses)	□	)
+ (plus)	&	+
- (hyphen) — (long dash)	@	-
= (equals)	#	=
$ (dollar sign)	$	$
/ (fraction mark) (and/or)	/	/
[(begin bracket)	*%	*(
] (close bracket)	*□	*)
' (apostrophe) (prime) (single quote)	**A	**A
% (per cent)	**K	**K
: (colon)	**C	**C
? (question mark)	**I	**I
∠ (angle)	**L	**L
X (multiplication sign)	**M	**M
" (quotes)	See Section 2.10	
; (semicolon)	**S	**S
° (degree or temperature) (angle sign)	*&0	*+0
↑ (direction sign)	**V	**V
↓ (direction sign)	**W	**W
! (exclamation point)	**X	**X
← (direction sign)	**&	**+
→ (direction sign)	**,	**,
÷ (division mark)	/	/
superscript	*&	*+
subscript	*@	*-

2.15 When a mark, symbol, or other situation is encountered which is not accommodated by any of the foregoing provisions, the patent text copy is marked in red and the symbol

**B

is punched to represent the omitted portion of the patent text.

2.16 When an error is made in the punched text, the card in which the error occurred is repunched. When the error includes an omission from the text, a card or cards are inserted to contain the overflow text.

When the last non-space of inserted text is in the middle of a text word, or when the total number of blanks between the last non-space on the insert card and the first non-space on the following card is *72 or greater*, the *cancel space* symbol

* space

is punched immediately after the inserted text. This symbol serves to cancel the * and all *spaces* after the * so that the text with the words and spacing shown in the document is preserved.

The procedure for correcting errors is illustrated in the following examples. The text from which the examples are taken reads as follows:

> Another object of the invention is to provide a piston pump capable of operating at high temperatures and having improved piston expansion characteristics to reduce loss of volumetric efficiency. More specifically, the invention aims to provide a pump having the ability to compensate . . .

Ex. 1--The following was punched:

```
1                                   72
|                                   |
.............................CAPABLE OF

OPERATING.....................COMP
```

Since the F of OF occurred in column 72, there should have been a space in column 1 of the next card before the word OPERATING. The error is corrected as follows:

```
1                                   72
|                                   |
.............................CAPABLE OF      (unchanged)

 *                                          (insert card)

OPERATING.....................COMP           (unchanged)
```

The insert card provides a space in column 1. Since this is the only omission, the * is punched in column 2 with no punches in the rest of the card. Therefore, only the space in column 1 of the insert card becomes part of the punched text.

Ex. 2--The following was punched:

```
1                                   72
|                                   |
ETRIC.....................*MRE SPECIFICA

LLY, THE INVENTION.............COMPENSA
```

The letter O was omitted from the word *MORE. The error is corrected as follows:

```
1                                       72
|                                       |
ETRIC.....................*MORE SPECIFIC        (new card)

A*                                              (insert card)

LLY, THE INVENTION.............COMPENSA        (unchanged)
```

A new card is punched with the typographical error corrected. Since the error involved an omission, the text which overflows the new card is punched into the insert card. Since the last letter punched into the insert card is in the *middle* of a text word, the * is punched immediately after the letter and no further punches are made in the insert card. The punched text now mirrors the document text.

Ex. 3--The following was punched:

```
1                                       72
|                                       |
LOSS OF................*MORE SPECIFICY,

THE INVENTION..................COMPENSA
```

The letters ALL were left out of the word SPECIFICALLY. The error is corrected as follows:

```
1                                       72
|                                       |
LOSS OF................*MORE SPECIFICALL      (new card)

Y,                                              (insert card)

THE INVENTION..................COMPENSA      (unchanged)
```

A new card is punched with the typographical error corrected. Since the error involved an omission, the text which overflows the new card is punched into the insert card. In comparison with Example 2 above, the last mark punched into the insert card in Example 3 is at the *end* of a text word. Since a "space" may consist of up to 71 blanks in a card [see Section 2.6] and only 70 blanks remain to the next non-space, no additional symbology is required to be punched into the insert card.

Ex. 4--The following was punched:

```
1                                       72
|                                       |
ETRIC.....................*MOORE SPECIFI

CALLY, THE INVENTION...........COMPENSA
```

An extra letter O was punched in the word *MORE. The error is corrected as follows:

```
1                                       72
|                                       |
ETRIC.....................*MORE SPECIFI      (new card)

CALLY, THE INVENTION...........COMPENSA      (unchanged)
```

Since a letter was *removed* from the punched text, an additional space to take the place of this letter is left in the new card. No insert card is necessary. In the above example, an additional space was left before the word *MORE. [See Section 2.6 on the definition of "space"].

Ex. 5--The following was punched:

```
1                                          72
|                                          |
*ANOTHER.................PISTON CAPBLE

 OF OPERATING.....................COMP
```

The word PUMP was omitted from the text and the letter A was omitted from the word CAPABLE. The error is corrected as follows:

```
1                                          72
|                                          |
*ANOTHER.................PISTON PUMP C      (new card)

APABLE                                     (insert card)

 OF OPERATING.....................COMP      (unchanged)
```

A new card replaces the erroneous one. The overflow text is punched into the insert card. Since the total number of spaces from the last non-space on the insert card (which is the *end* of a word) to the next non-space on the following card is less than 71, no additional symbology is required on the insert card.

2.17 The following symbols are specifically excluded:

#

&

Notations will be designated for these symbols if they are needed.

[See Section 2.15 for other symbols not provided for].

3.0 SYMBOLOGY PECULIAR TO PATENT TEXT

3.1 The following information is extracted from the patent heading in the following order:

(1) The patent number
(2) The full title of the invention
(3) The full name(s) of the inventor(s)

The patent number (1) needs no comment.

3.1.1 The title of the invention (2), while often printed in upper case on the patent, is punched entirely as if it were in lower case. This procedure applies to every letter in the title *including* the first letter of the first word of the title and the first letter of a proper name or geographic location in the title, e.g.:

Ex. 1--Control Mechanism for Pump

is punched

```
CONTROL MECHANISM FOR PUMP
```

Ex. 2--CONTROL MECHANISM FOR PUMP

is punched

```
CONTROL MECHANISM FOR PUMP
```

Ex. 3—Flow Through a Bernouilli Tube

is punched

```
FLOW THROUGH A BERNOUILLI TUBE
```

Ex. 4—Synthesis of Iranian-Type Oil

is punched

```
SYNTHESIS OF IRANIAN@TYPE OIL
```

3.1.2 When there is more than one inventor (3), the inventors' names are separated by periods.

When the inventor is deceased, the name(s) of his representative(s) and the representative(s)'s title (such as executor or administratrix) are included in the patent heading. These names are punched according to the following format: inventor's name followed by a **/ followed by the name of the representative. The representative's title ***must*** begin with a ***lower case*** letter.

When the inventor has changed his name, his new name is punched followed by **/ FORMERLY followed by his original name.

There are ***no spaces*** on either side of the **/ as follows:

Ex. 1—John Doe and James Roe; Mary Doe executrix of the estate of John Doe

is punched

```
*JOHN *DOE**/*MARY *DOE EXECUTRIX. *JAMES *ROE
```

Ex. 2—John Doe; Mary Roe, administratrix of the estate of James Roe

is punched

```
*JOHN *DOE. *JAMES *ROE**/*MARY *ROE ADMINISTRATRIX
```

Ex. 3—James Roe, now by change of name James T. Roe Co.

is punched

```
*JAMES *T. *ROE *CO.**/FORMERLY *JAMES *ROE
```

Ex. 4—Henry Phillips and William E. Hunt, deceased, by Josephine Hunt and Annie Boswell, executrices

is punched

```
1                                                                    72
|                                                                    |
*HENRY *PHILLIPS. *WILLIAM *E. *HUNT**/*JOSEPHINE *HUNT AND *ANNIE *BOSW

ELL EXECUTRICES
```

The ***assignment*** information, application date, and serial number are ***not*** copied.

3.1.3 The last character of each of the following items of information: (1) patent number, (2) title, and (3) inventors' names, etc., is separated from the first character of the next item of information by a ***minor division*** symbol

space / space

Thus, the heading of patent 2,765,622 which reads as follows:

> CONTROL MECHANISM FOR PUMP AND MOTOR FLUID SYSTEM. Don R. Hill and Ernest C. Chasser, Warren, Ohio; Martha H. Hill, executrix of said Don R. Hill, deceased. Application April 27, 1950, Serial No. 158,374.

is punched

```
1                                                                        72
|                                                                        |
2,765,622 / CONTROL MECHANISM FOR PUMP AND MOTOR FLUID SYSTEM / *DON *R.

*HILL**/*MARTHA *H. *HILL EXECUTRIX. *ERNEST *C. *CHASSER /
```

3.2 The beginning of each paragraph is indicated by a notation which furnishes the column number in which the paragraph occurs and the number of the line on which it begins. The column number is preceded by a * and is followed by a @ (a hyphen on the Fortran Punch) and the line number, e.g.:

a paragraph beginning in column 2 at
line 47 with the word The is punched

```
*2@47 *THE
```

3.3 Formulas (chemical or mathematical), tabular material, and diagrams which appear in the patent text are not copied. When such material is encountered, the ***begin formula*** symbol

**F

is punched and nothing more is punched in the remainder of the card. ***The formula, table, or diagram is encircled in red on the patent text.*** The next card is begun with the ***end formula*** symbol

**G space

and punching of the patent text is continued. For example, lines 75 through 81 of column 4 of 2,706,891 which read as follows:

> of such end 116. Assuming that the area of end 116 is, say, seven times the area of the end 118 of piston 112, by the formula
>
> $$P=\frac{F}{A}$$
>
> the pressure developed on the fluid in the chamber 119

are punched

```
1                                                                        72
|                                                                        |
OF SUCH END 116. *ASSUMING THAT THE AREA OF END 116 IS, SAY, SEVEN TIMES

 THE AREA OF THE END 118 OF PISTON 112, BY THE FORMULA **F

**G THE PRESSURE DEVELOPED ON THE FLUID IN THE CHAMBER 119
```

Ending the punched text on a card with the begin formula symbol will facilitate insertion of the formula, table, or diagram at a future time.

3.4 When a subtitle occurs in a column, the subtitle is introduced by the ***begin subtitle*** symbol

**N

and the subtitle is copied in *lower case* letters regardless of the kind of type in which the subtitle appears in the text. The subtitle is followed by the *end subtitle* symbol

**P

For example, the following lines from column 3 of 2,706,891

actuation thereof.

Operation

line 9 Assuming that the plunger 163 of the hydraulic press 133 is in the down position,

are punched

```
1                                                                        72
|                                                                        |
ACTUATION THEREOF. **NOPERATION**P *3@9 *ASSUMING THAT THE PLUNGER 163 O

F THE HYDRAULIC PRESS 133 IS IN THE DOWN POSITION,
```

3.5 Claims, being paragraphs, are introduced with the paragraph notation (see 3.4 above) before the claim number. The period after the claim number is treated as a period after an abbreviation, e.g.:

claim 4 which begins in column 6 at line 29

is punched

```
*6@29 4**. *THE HYDR...
```

3.6 At the end of the claims, the *begin references cited* symbol

**D

is punched. The words "References cited in the file of this patent" and similar words are not punched. The references are punched in the order in which they appear in the patent.

A single virgule / is used as a *minor division* for separating individual references from each other. Two consecutive virgules // are used as a *major division* for separating groups of references. Application of these notations is amplified below.

The *United States* patents are punched as follows:

number space inventor space date/number space inventor space date/number ...../...etc...

The *major division* symbol // follows the last character punched for the last United States patent.

The *foreign* patents are punched as follows:

number space country space date/number space country space date/number ...../...etc...

The *major division* symbol // follows the last character punched for the last foreign patent.

Other reference material is punched as though it were regular patent text. When there is more than one such reference, one virgule / is punched at the end of each reference *except* the last of such references.

When no U. S. patents are references, two virgules are punched after the **D and precede the first foreign patent, as follows:

**D//

When no U. S. or foreign patents are references, two virgules are punched to signify the absence of *each* group of patents and precede any other reference material, as follows:

**D////

After all of the references are punched, the *end references cited* symbol

**E

is punched.

When no references are cited, the begin references cited symbol followed by four virgules followed by the end references cited symbol are punched, as follows:

D////E

Thus, the following text

References Cited in the file of this patent

UNITED STATES PATENTS

Number	Name	Date
2,234,215	Youker	Mar. 11, 1941
2,376,350	Fryling	May 22, 1945
2,469,017	Sundet	May 3, 1949

OTHER REFERENCES

India Rubber World, July 1949, page 476.
Kluchesky et al., Ind. and Eng. Cheml., vol. 41, No. 8, August 1949, pp. 1768-1770.

is punched

```
1                                                                      72
|                                                                      |
**D2,234,215 *YOUKER *MAR**. 11, 1941/2,376,350 *FRYLING *MAY 22, 1945/2

,469,017 *SUNDET *MAY 3, 1949////*INDIA *RUBBER *WORLD, *JULY 1949, PAGE

476./*KLUCHESKY ET AL**., *IND**. AND *ENG**. *CHEML**., VOL**. 41, *NO*

*. 8, *AUGUST 1949, PP**. 1768@1770.**E
```

3.7 If a patent has a Certificate of Correction, the text of the patent is first corrected by hand according to the Certificate, and the text is then punched as though it had been originally issued without the error which the Certificate corrected.

3.8 After each patent, one blank card is inserted.
[See Section 2.2]

3.9 No provisions are presently included for punching Reissue patents or Reissues of Reissue patents. The following portions of patent text are specifically excluded from present consideration:

Drawings
Preamble
Application number and date
Assignment information
Classification
Footnotes

Symbology and instructions for these areas will be introduced subsequently if needed. [See Section 3.3 with respect to formulas].

4.0 SYMBOL DICTIONARY

Table 5 is a composite of (a) several sets of symbols which are expedient for use as identifying devices and (b) equivalents for those of the symbols which have been selected for use in the current patent text punching project.

The symbols are given in the standard IBM 024 and Fortran Punch characters. The meanings for the symbols which have been used in this manual are given in the third column. Reference is made in the right-hand column to the Section and Subsection in which the symbol is discussed in the foregoing part of this report.

The remaining symbols or other synthesized sets of symbols will be used if needed for data not presently provided for herein.

TABLE 5.–SYMBOL DICTIONARY

IBM 704-FORTRAN SYMBOL	STANDARD IBM 024 SYMBOL	SYMBOL MEANING	REFERENCE
NO PUNCH			
space	space	space between words	2.6
blank card	blank card	end of document	2.6;3.8
SINGLE UNITS			
1	1	1	
2	2	2	
3	3	3	
.	.	. Arabic numerals	2.4
.	.	.	
.	.	.	
9	9	9	
0	0	0	
A	A	a	
B	B	b	
C	C	c	
.	.	. lower case letters	2.3
.	.	.	
.	.	.	
Z	Z	z	
. space	.	end of sentence	2.7
.	.	decimal point	2.7
–	@	minus; hyphen; long dash	2.14
(	%	(start parentheses	2.14
)	¤	) close parentheses	2.14
/	/	in fractions; in and/or;	2.14
		÷ division mark; separates references	3.6

TABLE 5.--SYMBOL DICTIONARY--Con.

IBM 704-FORTRAN SYMBOL	STANDARD IBM 024 SYMBOL	SYMBOL MEANING	REFERENCE
SINGLE UNITS--Con.			
space / space	space / space	separates items in patent heading	3.1
=	#	= equals	2.14
,	,	, comma	2.14
$	$	$ dollar sign	2.14
+	&	+ plus	2.14
PRECEDED BY A SINGLE STAR			
* space	* space	cancels * and all space to the next non-space	2.16
*A	*A	A	
*B	*B	B	
*C	*C	C	
.	.	. upper case letters	2.3
.	.	.	
.	.	.	
*Z	*Z	Z	
*1	*1		
*2	*2		
*3	*3		
.	.	not to be used	
.	.		
.	.		
*9	*9		
*0	*0		
* col. no. - line no. space	* col.no. @ line no. space		
e.g.:	e.g.:		
*1-1 space	*1@1 space		
*1-15 space	*1@15 space	beginning of paragraph; the number after the * is the column number and the number after the @ is the number of the line on which the paragraph begins.	
*1-26 space	*1@26 space		
.	.		3.2
.	.		
.	.		
*2-4 space	*2@4 space		
.	.		
.	.		
.	.		
*24-65 space	*24@65 space		
*-	*@	subscript	2.12
*(	*%	[begin bracket	2.14
*)	*□	] close bracket	2.14

TABLE 5.--SYMBOL DICTIONARY--Con.

IBM 704-FORTRAN SYMBOL	STANDARD IBM 024 SYMBOL	SYMBOL MEANING	REFERENCE
PRECEDED BY A SINGLE STAR--Con.			
*/	*/	begin Roman numeral	2.5
*=	*#	begin italics or underscoring	2.9
*,	*,	end Roman numeral	2.5
*$	*$	end italics or underscoring	2.9
*+	*&	superscript	2.12
*.	*.	not to be used	
PRECEDED BY TWO STARS			
**A	**A	prime; apostrophe; single quote	2.11
**B	**B	mark for which no provision made	2.15
**C	**C	: colon	2.14
**D	**D	begin references cited	3.6
**E	**E	end references cited	3.6
**F	**F	begin formula	3.3
**G space	**G space	end formula	3.3
**H	**H	... or *** ellipsis	2.7.6
**I	**I	? question mark	2.14
**J	**J	(available for later use)	
**K	**K	% per cent sign	2.14
**L	**L	∠ angle mark	2.14
**M	**M	X multiplication sign	2.14
**N	**N	begin subtitle	3.4
**O	**O	not to be used	
**P	**P	end subtitle	3.4
**Q	**Q	" begin quotation	2.10
**R	**R	(available for later use)	
**S	**S	; semicolon	2.14
**T	**T	(available for later use)	
**U	**U	" end quotation	2.10
**V	**V	↓ direction sign down	2.14
**W	**W	↑ direction sign up	2.14
**X	**X	! exclamation point	2.14
**Y	**Y	lower case Greek letter	2.13
**Z	**Z	upper case Greek letter	2.13
**-	**@	(available for later use)	

TABLE 5--SYMBOL DICTIONARY--Con.

IBM 704-FORTRAN SYMBOL	STANDARD IBM 024 SYMBOL	SYMBOL MEANING	REFERENCE
PRECEDED BY TWO STARS--Con.			
**(	**%	begin capitalized word(s)	2.3
**)	**¤	end capitalized word(s)	2.3
**/	**/	separates the name of the inventor from the name of his representative	3.1.2
**=	**#	begin bold-face type	2.8
**,	**,	→ direction sign right	2.14
**$	**$	end bold-face type	2.8
**+	**&	← direction sign left	2.14
**.	**.	end of abbreviation; period after claim number	2.7.3; 3.5
**..	**..	end of abbreviation which is end of sentence	2.7.4
MISCELLANEOUS			
*+0	*&0	° degree or temperature mark; angle mark	2.12
//	//	end of U. S. patent references; no U. S. patent references cited	3.6
////	////	no U. S. and no foreign patent references cited	3.6
AA	**A**A	'' double prime	2.11
AA**A	**A**A**A	''' tripe prime	2.11

5.0 THE IBM 704 PROGRAMS

5.1 Two programs nave been prepared at the Massachusetts Institute of Technology. The first program takes the text from the punched cards, closes up the excess spaces which were introduced, and transliterates this material into a standardized output format on an IBM 704 tape. The second, or search, program locates, in context, any desired *text* word or words in this tape and reproduces it (them) close to the center of approximately 100 *machine* words, in a format suitable for mounting on a 3 x 5 card.

5.2 The 36 binary digit (bit) word of the IBM 704 is divided for the present project into six Binary Coded Decimal (BCD) characters. The most convenient arithmetic operations are performed on entire machine words. This suggests that, for a high-speed search for a text word or a group of text words, the text words should first be arranged in a regular format with respect to the beginnings and ends of machine words.

5.3 The following machine word format is used:

Each text word begins in the *second* BCD position in a machine word, the first position being filled with a blank. If the text word is more than 5 letters in length, it continues into the next machine word, the sixth letter occupying the first BCD position and the remaining letters following consecutively. A text word is terminated by filling the remaining positions of the last machine word into which it extends, if any such positions remain, with blanks.

Numbers, in general, have the same word format as text words.

Exception: The patent number and the column-line paragraph notation. These are exemplified in Section 5.5 below.

Sample output word formats are presented in Table 6 together with their text and Fortran equivalents.

Table 6.--Output Word Formats

TEXT	IBM 704-FORTRAN SYMBOL	MACHINE WORD BCD Positions					
		1	2	3	4	5	6
this	THIS		T	H	I	S	
apertures	APERTURES	U	A R	P E	E S	R	T
aperture near	APERTURE NEAR	U	A R N	P E E	E A	R R	T
valve 101	VALVE 101		V 1	A 0	L 1	V	E

5.4 Punctuation marks and other typographic symbols are subdivided into four categories.

5.4.1 Each of the first group of punctuation marks and other symbols ***occupies an entire machine word.*** This group is composed of two types of notations:

(a) Simple punctuation marks. The machine word for each of these marks begins in the third BCD position. The word formats for this group are shown in Table 7.

Table 7.--Word Formats

Symbols which occupy an entire machine word:

(a) Simple punctuation marks

IBM 704-FORTRAN SYMBOL	MACHINE WORD BCD Positions						SYMBOL MEANING
	1	2	3	4	5	6	
. space			.				period (end of sentence)
**H			.	.	.		ellipsis
**X			.	X			exclamation point
**I			.	Q			question mark
,			,				comma
**S			,	S			semicolon
**C			,	C			colon
**K			0	/	0		per cent
/			/				virgule (minor division)
//			/	/			two virgules (major division)

(b) Pairs of punctuation marks and other symbols for which there are complementary notations for the commencement and the termination of the condition represented by the symbol pairs. The machine word for the "begin" mark of the pair starts in the fifth BCD position. The "end" mark of the pair begins with a) in the third BCD position.

Exception: The "begin parentheses" machine word occupies only the sixth BCD position.

The word formats for this group are shown in Table 8.

Table 8.--Word Formats

Symbols which occupy an entire machine word:

(b) Pairs of punctuation marks

IBM 704-FORTRAN SYMBOL	MACHINE WORD BCD Positions						SYMBOL MEANING	
	1	2	3	4	5	6		
( )			)			(	begin end	parentheses
*= *$			)	 I	I	(	begin end	italics
**Q **U			)	 Q	Q	(	begin end	quotes
**= **$			)	 B	B	(	begin end	bold face
**(**)			 (	 C	C	(	begin end	capitalized word(s)
*/ *,			)	 N	N	(	begin end	Roman numerals
**N **P			)	 S	S	(	begin end	subtitle
**F **G			)	 F	F	(	begin end	formula
*(*)			)	 K	K	(	begin end	brackets
**D **E			)	 R	R	(	begin end	references cited
			)	 E	E	(	detected error isolating symbol; see Section 5.8	

5.4.2 Each mark of the second group is *part of a machine word.* The BCD position in which the mark is noted in the output tape depends on the text word in which the mark is contained. The marks of this group are the

abbreviation period

period non-space (e.g., a decimal point in a number: 1.259)

comma non-space (e.g., in a number: 1,259)

apostrophe

In the output, both periods appear as periods. However, both the comma and the apostrophe appear as commas.

Examples of machine words containing the symbology for these marks and the corresponding text words are given in Table 9.

Table 9.--Word Formats

Periods and Commas

TEXT	IBM 704-FORTRAN SYMBOL	MACHINE WORD BCD Positions					
		1	2	3	4	5	6
etc.	ETC**.		E	T	C	.	
3,176.45	3,176.45		3 4	, 5	1	7	6
76.45	76.45		7	6	.	4	5
an arm's length	AN ARM**AS LENGTH	H	A A L	N R E	M N	, G	S T

5.4.3 Two symbols which occupy entire machine words but differ because of word format or substance from the foregoing two groups are given in Table 10.

5.4.4 The last category is a residual one for all other marks and symbols. There is no transliteration for these marks; they appear in the machine word output in IBM 704-Fortran symbology. They occupy *parts of machine words* in the same manner as the marks of group (2) when they are parts of words in patent text. They start *new machine words* which begin in the second BCD position only when they are preceded by a space in patent text. The series of symbols which constitute a single notation appear in consecutive positions in a machine word rather than in new machine words. Illustrative word formats are shown in Table 11.

Table 10.--Word Formats

Special Group

IBM 704-FORTRAN SYMBOL	MACHINE WORD BCD Positions						SYMBOL MEANING
	1	2	3	4	5	6	
*						*	letter following is upper case
**B			*	U	S		undefined symbol

Table 11.—Word Formats

Illustrations for All Other Marks

TEXT	IBM 704-FORTRAN SYMBOL	MACHINE WORD BCD Positions					
		1	2	3	4	5	6
450°	450*+0	0	4	5	0	*	+
∠ abc	**L ABC		* A	* B	L C		

5.5 The patent number and column-line paragraph notation always occur together. In transliterated form, the patent number follows a (which appears in the third BCD position of the machine word; remaining spaces of the second machine word into which the number extends, if any such spaces remain, are filled with blanks. The column-line paragraph notation starts in the fifth BCD position and is followed by a) and the remaining spaces of the machine word are filled with blanks. The program for creating the IBM 704 tape inserts the patent number and paragraph notation at intervals of about 120 machine words in long paragraphs of patent text. The repeated number differs from its first occurrence by the addition of a + in the fourth BCD position immediately preceding the paragraph notation.

The machine words for patent number and paragraph notation for patent 2,709,339 and paragraph beginning in column 2 at line 51 are shown in Table 12.

Table 12.—Word Formats

Patent and Paragraph Numbers

	MACHINE WORD BCD Positions					
	1	2	3	4	5	6
First occurrence	0 5	9 1	(,)	2 3	, 3 2	7 9 -
Repeated occurrence	0 5	9 1	(,)	2 3 +	, 3 2	7 9 -

5.6 A graphic description of the machine-word structure of the transliterated text, as it is stored on magnetic tape, is given in Figure 1. In the description, the characters are replaced by BCD codes and the direction of the tape is toward the bottom of the page. The description is for the following text from patent 2,709,339 which begins a paragraph starting in column 2 at line 66:

> The present invention contemplates a PUMP capable of developing *pressures* as high as 5000 p.s.i., while operating (with a suitable fluid) at temperatures up to 450° Fahrenheit; the two-stage pumping system is shown in Fig. 2' in which D at point 36*a* stands for "LOAD."

BCD Positions					
1	2	3	4	5	6
		(	2	,	7
0	9	,	3	3	9
				2	-
6	6	)			
					*
	T	H	E		
	P	R	E	S	E
N	T				
	I	N	V	E	N
T	I	O	N		
	C	O	N	T	E
M	P	L	A	T	E
S					
	A				
				C	(
	P	U	M	P	
		)	C		
	C	A	P	A	B
L	E				
	O	F			
	D	E	V	E	L
O	P	I	N	G	
				I	(
	P	R	E	S	S
U	R	E	S		
		)	I		
	A	S	G	H	
	H	I			
	A	S	O	O	
	5	O			
	P	.	S	.	I
.					
		,			
	W	H	I	L	E
	O	P	E	R	A
T	I	N	G		
		I	T	H	(
	W				
	A				
	S	U	I	T	A
B	L	E			
	F	L	U	I	D
		)			
	A	T			
	T	E	M	P	E

BCD Positions--Con.					
1	2	3	4	5	6
R	A	T	U	R	E
S					
	U	P			
	T	O			
	4	5	O	*	+
O					*
	F	A	H	R	E
N	H	E	I	T	
		,	S		
	T	H	E		
	T	W	O	-	S
T	A	G	E		
	P	U	M	P	I
N	G				
	S	Y	S	T	E
M					
	I	S			
	S	H	O	W	N
	I	N			
					*
	F	I	G	.	
	2	,			
	I	N			
	W	H	I	C	H
					*
	D	T			
	A	O	I	N	T
	P	6			
	3			I	(
	A				
		)	I		
	S	T	A	N	D
S					
	F	O	R	Q	(
				C	(
	L	O	A	D	
		)	C		
		)	Q		
		.			

GRAPHIC DESCRIPTION OF TRANSLITERATED TEXT

Figure 1

5.7 An example is shown in Figure 2 of an output format of the search program for the phrase

aircraft landing gear

The program provides a print-out of 48 columns (8 machine words) and 16 lines, with the phrase near the middle, so that the phrase can be read in context. The ninth line begins with a mnemonic of the phrase, in this case

ALG

immediately preceded by an * and followed by a) and then the phrase itself.

MACHINE WORDS

	1	2	3	4	5	6	7	8
	*	THE	PILOT	UNIT	*	E	IS	PRESS
	URE	SENSI	TIVE	,	RESPO	NDING	TO	THE
	OUTPUT		PRESS	URE	OF	PUMP	*	C
	,	WHICH	WILL	TEND	TO	VARY	WITH	CHANG
	ES	IN	(2,709,339		+2-51)		THE	LOAD
	DEMAND		.	*	FOR	EXAMPLE		,
	ASSUMING		THAT	THE	LOAD	COMPRISES		THE
	HYDRAULIC		ACTUA	TING	CYLIN	DERS	OF	AN
Phrase searched →	*ALG)	AIRCR	AFT	LANDI	NG	GEAR	,	CONNE
	CTED	TO	THE	PUMPI	NG	SYSTEM		BY
	A	SUITA	BLE	CONTR	OL	VALVE	(	NOT
	SHOWN	)	WHICH	IS	OPENED		WHEN	THE
	PILOT	DESIRES		TO	ACTUATE		THE	LANDI
	NG	GEAR	,	THE	NORMAL		CONDI	TION
	WILL	BE	ONE	IN	WHICH	THE	VALVE	IS
	CLOSED		,	CAUSI	NG	THE	PRESS	URE

OUTPUT FORMAT OF SEARCH PROGRAM

Figure 2

5.8 The program for transliterating the punched card "alphabet" to the IBM 704 tape symbology also detects and displays errors in punching. The detected error isolating symbols are shown in Table 8.

The display begins with the symbol

E(

which is consecutively followed by

(1) a description of the type of error

(2) three blocks of five characters each in BCD positions 2 through 6, each block preceded by a virgule / in the first position

(3) a virgule / in the first position of the following machine word and the symbol)E in positions 3 and 4.

A sample display on detection of the improper punching of patent number 2,364,357 as 2,364K357 is illustrated in Figure 3.

MACHINE WORDS

1	2	3	4	5	6	7	8
123456	123456	123456	123456	123456	123456	123456	123456
E(	IMPROPE	R PAT	. NO.	/	/ 2,3	/64K35	/)E

SAMPLE ERROR DISPLAY

Figure 3

6.0 BIBLIOGRAPHY

ACM-GAMM Committee [Association for Computing Machinery (U. S.)--Association for Applied Mathematics and Mechanics (Germany)]. "Preliminary Report--International Algebraic Language." *Communications of the Association for Computing Machinery* 1:12 (Dec. 1958) 8-22.

Edmundson, H. P., D. G. Hays, and R. I. Sutton. *Studies in Machine Translation--3: Resume of Machine Codes and Card Formats.* Project Rand Research Memorandum RM-2064. Santa Monica, Rand Corp., 1958. (ASTIA no. AD 156048). Formats and codes for punching textual, glossary, and translation data are presented.

Edmundson, H. P., K. E. Harper, D. G. Hays, and A. M. Koutsoudas. *Studies in Machine Translation--9: Bibliography of Russian Scientific Articles.* Project Rand Research Memorandum RM-2069. Santa Monica, Rand Corp., 1958. (ASTIA no. AD 210147). Eight corpora of Russian scientific papers (227,752 running words) are described.

Luhn, H. P. *An Experiment in Auto-Abstracting: Auto-Abstracts of Area 5 Conference Papers, International Conference on Scientific Information, Washington, D. C., November 16-21, 1958.* Yorktown Heights, N. Y., IBM Research Center, 1958. Representations of output punctuation marks not available in the Hollerith code are listed in the Explanatory Notes Section.

Luhn, H. P. "General Rules for Creating Machinable Records for Libraries and Special Reference Files." Excerpt from *Information Storage and Retrieval Installations.* Yorktown Heights, N. Y., IBM Research Center, 1957, rev. 1958.

Secrest, B. W. *The IBM Electronic Statistical Machine Applied to Word Analysis of the Dead Sea Scrolls.* New York, IBM World Trade Corp., 1958. Pages 2 and 3 contain output symbols which, in this case, are the same as the punching codes.