LIBRARY

FOR

MECHANICAL

TRANSLATION

using

punched card machinery.

K.Spärck Jones
R.M.Blackmore
C.Wordley

Cambridge Language Research Unit.

INPUT CARDS. I.PACK.

Stage 1. Production of the chunked text by means of the Chunking Reference Dictionary.

The input text is chunked, i.e. the words are split into smaller units, using a permanent Reference Dictionary. This lists the words in alphabetical order, giving the way in which they are chunked. The number of possible different meanings is noted against each entry, e.g. -a = 8. Words are usually chunked into stems and endings, for although chunking is done for reasons of economy, it is useful to have chunks with syntactic or semantic meaning in their own right.
E.G. GALLI/A/ EST/ OMN/IS/ DIVIS/A/ IN/ PART/IS/ TR/ES/ ;/
One mechanical procedure for chunking is envisaged by Richens and Halliday in "Word Decomposition for Machine Translation". An improved method is described by R.M.Needham "Problems of Chunking".
Each chunk, or meaning of a chunk, is represented by an input card.

Stage 2. Text Position Indicators and Serial Numbers.

To enable the chunks to be rearranged in their original order whenever necessary, each chunk is given a Text Position Indicator showing its place in the word, and the position of the word in the sentence and paragraph.

E.G.(Para 1, GALLI/A/ EST/ OMN/IS/ DIVIS/A/ IN/ PART/IS/ TR/ES/ ;/
Sent 1.) 1.1 1.2 2. 3.1 3.2 4.1 4.2 5. 6.1 6.2 7.1 7.2 8.
This is given on the input card as follows: positions 31-51 in row X are used, divided into four sections; 31-36 (paragraph) 37-42 (sentence) 43-48 (word) 49-51 (chunk). The numbers are coded in binary with the following result:
E.G. OMN/IS/: OMN/ 00000100000100001100 1
IS/ ..... 00000100000100001101 0.

Each 1 occurring in the coding indicates a hole punched in the card.
Each chunk in the text is also serially numbered, for reference purposes and this number is stamped onto the card,
The chunk and its English equivalent is also written on the card.

Stage 3. Coding of the chunk spelling.

The most convenient method of accommodating the spelling on Hollerith cards is as follows:
Alphabetic dictionary entries are in 20 bits which are punched in columns 79 and 80.
Arrange the letters as in the example below, separating the initial letter from the rest. (As the coding of the other letters may involve conflation this is to ensure that a dictionary reference can always be made to the card.)

E.G. C A E S V I R G
  A R I L I
  A

Look up in the table the symbols for each letter and arrange in the same way.

E.G. C A E S
  00011 00001 00101 10011
  A R
  00001 10010

Obtain a row of 20 bits by adding up the columns: if there are only 0's write 0, if an even number of 1's write 0, if an odd number of 1's write 1.

E.G. Caesar: 00011000001011110011.

Divided into two, this result is punched vertically in columns 79 and 80, a 1 being indicated by a hole.

CODE

A 00001	I 01001	Q 10001	Y 11001
B 00010	J 01010	R 10010	Z 11010
C 00011	K 01011	S 10011	, 11011
D 00100	L 01100	T 10100	. 11100
E 00101	M 01101	U 10101	! 11101
F 00110	N 01110	V 10110	? 11110
G 00111	0 01111	W 10111	: 11111
H 01000	P 10000	X 11000	; 00000

A preceding hyphen is indicated in position 79 in row X, a following in position 80 in row X.

Stage 4. Sorting the input pack into alphabetical order preparatory to Dictionary matching.

Owing to the compressed alphabetical coding it is only possible to sort the chunks by their first letters. As there are 30 symbols for letters or punctuation signs 29 sorts are required.

lst sort. Each symbolfor the first letter having 5 bits, sorting by the last bit first, which may be either a 1 or an 0, divides the cards into two packs. Each letter being taken as a number, this gives

1 3 5 7 9 11 13 15 17 19 21 23 25 27 29
2 4 6 8 10 12 14 16 18 20 22 24 26 28 30

2nd sort

These are in turn sorted using the next bit; for the first pack above this gives

1 5 9 13 17 21 25 29
3 7 11 15 19 23 27

3rd sort. The four packs obtained are then sorted by the next bit; the first pack given by the second sort divides thus

1 9 17 25 5 13 21 29

4th sort. The eight packs generated are then sorted by the fourth bit, giving, for the first pack in the previous sort

1 17 9 25

5th sort. Some of the packs given by the last sort will have only one card in them. Those which have two must be sorted by the fifth bit. This will give, for the first pack in the previous sort

1 and 17.
The sequence of cards finally obtained is
1 17 9 25 5 21 13 29 3 19 11 27 7 23 15 2 18 10 26 6 22 14 30 4 20 12 28 8 24 16.
The letters having been separated, the sequence must now be alphabetically re-ordered; this is best done by hand.
In each stage of sorting, the number of packs obtained is always double that of the previous sort, except when the penultimate stage generates some packs with only one member. This last will always be the case except when the last letter, defined as a number, is a member of the geometrical progression 1 2 4 8 16 etc.
When the alphabetic reordering has been done, if there is only one card for any letter, the pack is then in alphabetical order, If there is more than one card for any letter, these, due to the compressed spelling, must be reordered by hand using the written spelling of the chunk.

DICTIONARY

Stage 5. Dictionary matching.

This is the procedure for obtaining thesauric and syntactic information relevant to the chunks of the input text, this information being given in interlingual terms. The Dictionary is a permanent set of cards, there being one for each chunk or meaning of a chunk. The cards are selected from the Dictionary by matching using the coded spelling of the input cards. As both sets of cards are arranged alphabetically, a single collation is sufficient. When, however, a chunk has more than one meaning, each being represented by a separate card only one of these cards would be drawn out. To avoid this it is necessary to produce the requisite number of input cards for any such chunks so that the dictionary cards for each meaning can be obtained.
The dictionary information concerning any chunk is as follows:
a) Monolingual information - giving a brief analysis of the main characteristics of the chunk in terms of the distinctive grammatical categories of the input language. (This is changed for each input language, being the only non-interlingual information given on the card.) Positions 11-31 in row X, each having a one-to-one correspondence with a list of monolingual categories. For Latin Declension, Conjugation, Gender and Number are indicated.
b) Interlingual information-
1)List numbers. A word with a technical or very restricted use is placed on a list concerning a particular topic, as say, "oxygen" on a list of chemical elements. A list may have up to 1024 items, each entry having a serial number which is given in binary in positions 1-10 in row X. Thus list number 57, representing, for example "tomato" in a list of vegetables becomes 0000111001.
2)Syntactic properties. The syntactic features of a chunk are analysed in terms of a list of 78 properties. Those properties expressed by a chunk are indicated by a set of holes in any positions 1-78 in row Y, there being a correspondence between the positions and the items on the list. (The use of lists enables a direct matching procedure to be used for unambiguous words at the output stage.)
3)Thesaurus heads. The range of meaning of a chunk is represented by a set of thesaurus heads. The thesaurus heads used are those given in Roget's Thesaurus, the number having been reduced from 1000 to 780 by

compacting. (See M.Shaw "Compacting Roget's Thesaurus".) The heads take positions 1-78 in rows 0-9.
E.G. "sap" is given by the heads

Intrinsicality	15
Destruction	140
Concavity	223
Fluidity	294
Deterioration	497

DICTIONARY CARDS. D PACK.

Stage 6. Reproduction of relevant dictionary information.
These cards contain the information which is on both the input cards and in the Dictionary, thus providing all the data for the interlingual stages of the programme. The cards selected from the Dictionary by matching with the input cards are reproduced, and the information on the input cards is then added to them; this may alternatively be described as making join cards. The original cards from the Dictionary are then restored.

Stage 7. Restoring the dictionary cards to text order.
This is best done by hand; sorting by every possible combination of the binary bits representing the paragraph, sentence word and chunk respectively which together make up the Text Position Indicator, would be an intolerably lengthy procedure.

Stage 8. Pun Removal.
The dictionary pack being in text order, it is possible to remove some redundant information: certain input chunks may be represented by more than one card, due to their having several distinct uses.
E.G. The chunk -a would have
1) 1st conjugation imperative sing.
2) 1st conjugation past part. plu.
3) Nominative sing, 1st declension.
4) Ablative sing, 1st declension.
5) Nominative plu.neut. all decl..
6) Accusative plu.neut.all dec1.

As many as possible of the irrelevant meanings, or "puns" are removed as follows; the chunk cards for a word are grouped together using the Text Position Indicators. Those words having more than one card for any component chunk are then selected. (This is probably easiest by hand, but could be done on a sorter). An intersection procedure is then carried out within the word; that is, the pun cards are matched with these of the other chunks, using the monolingual information only. In the test language, Latin, any chunks which do not intersect in all of the following groups a) conjugation/declension, b)gender and c)number, are rejected. This routine removes most, though not all, puns.

MATE DICTIONARY
Stage 9. Selection of mate cards.
The original Dictionary indicated, for each chunk, only those syntactic and semantic heads which the chunk itself could convey. Every chunk

has also a corresponding Mate Dictionary card which extends the meaning of the chunk by giving the syntactic and semantic contexts in which the chunk has a distinctive character based on combinations with other chunks. The thesaurus part of a Mate Dictionary card contains groups of related heads of which only one or a few members may be indicated by the chunk itself, or contains heads which refer to contexts in which the chunk in combination can figure.

E.G. The Latin chunk in will have the head Ingress 260 on its dictionary card. In combination with "aratro" it can figure in the distinctive phrase "in aratro" meaning "to the altar" The head Marriage 694 will be given, not by the Dictionary, but by the Mate Dictionary.

The thesaurus part of the dictionary card for a chunk gives all possible uses of the chunk; in the particular input text it is unlikely that more than a few of these will be represented, so that it is necessary to select these and remove irrelevant ones. The selection is obtained by an intersection procedure. This is made possible by the fact that in most pieces of discourse a particular context is emphasised by the repetition of relevant ideas in several words; the intersection procedure relies on this, as it selects only those heads which occur more than twice. It is however, sometimes the case that a distinctive concept is only directly indicated once, so that it would be lost by an ordinary intersection. To avoid this the intersection is extended to the information given in the Mate Dictionary, it being very unlikely that any concept relevant to the context of the piece of discourse would not be given in any of the Mate Dictionary cards for all the words in the passage. The procedure is as follows: using the coded spelling, the dictionary cards are matched with the Mate Dictionary and relevant cards drawn out.

MATE CARDS, M. PACK
Stage 10. Reproduction of Mate Dictionary cards.

The cards selected from the Mate Dictionary are reproduced and the originals restored.

SECOND MATE CARDS. M 2 PACK.
Stage 11. Mate card intersection procedure.

The mate cards are intersected with each other as follows: the first card in the mate pack is intersected with each of the others in turn and any intersecting holes are reproduced on a single card, the second mate card of the chunk in question. The second card is then intersected with all the pack except the first card and another second mate card generated. The third mate card is then intersected with the pack, the first two cards, and so on. As most mate cards indicate a large number of heads, the second mate cards will probably contain more heads than the original dictionary

cards; heads irrelevant to the context will, however have been eliminated.

THESAURUS CARDS. T PACK.

Stage 12. Intersection of second mate cards with dictionary cards.

The second mate cards are intersected with their corresponding dictionary cards in order to eliminate unwanted heads in the latter; the cards generated by this meet procedure, which contain the selected thesaurus heads, form the thesaurus pack.

WORD CARDS, W PACK.

Stage 13. Combination of chunks within the word.

Subsequent procedures are based on the whole word, so that the separation into chunks is now redundant. Using the Text Position Indicators the chunks within the words are combined and new cards, containing the information on all component chunk cards, are generated, one for each word. The word cards represent the join of the constituent chunk cards,// (The coded spelling is omitted from these cards.)

HEAD FREQUENCY SERIES OF CARDS

Stage 14. Selection of most frequently occurring heads within a predetermined length of text.

The context structure of a particular piece of discourse is indicated by the presence of a few frequently occurring heads. In order that these may be found, so that the context structure lattice scan be generated for the text the sentences in the text, the following procedure is carried out; they are based on these mathematical principles:

From an ordered pack of cards P_j, a second pack of cards P_{j+1}, is generated by the following method: for each card C_j in the pack P_j a card is produced on which there are holes where there are holes on the C_j card and on subsequent cards of the P_j pack. Let these cards, the C_{j+1} cards, in the order in which they were produced, be the P_{j+1} pack. It follows that, if a hole is punched n times in the P_j pack, it will be punched n-1 times in the P_{j+1} pack. For the last card occurring in the P_j pack will not give rise to an occurrence in the P_{j+1} pack. Therefore, if the first pack is P_1 any hole occurring at least twice will be punched on some card in the second pack, P_2. If a join card is produced for all the cards in the P_2 pack, this card will give the total set of holes occurring at least twice in the P_1 pack. Similarly, (by induction) the set of holes occurring at least r times may be obtained by the join of all the cards in the pack P_r.

If the sentence under consideration is GALLIA EST OMNIS DIVISA IN PARTIS TRES; the procedure would be thus: the word cards, in text order, constitute the initial pack for the above procedure. The top card in the pack is intersected with each of the following cards in turn, and all common holes punched on one card. The same procedure is then followed with each of the word cards in turn. The resulting pack can then be taken as the

initial pack for the next series of intersections; the procedure is followed until those heads occurring a determined number of times (or more) are left. (The pack of word cards is stored for future use). (Ref: C.Wordley "A Mechanical Procedure for finding the number of occurrences of Thesauric Head holes, over a pre-judged section of input text, thus enabling Sentence Structure Lattices to be constructed".) The most frequently occurring heads are then punched onto a grid card. The pack of word cards is then sorted by this grid so that the heads are assigned to the words which originally generated them. For convenience in the construction of the lattice these heads are indicated by letters rather than numbers.

CONTEXT STRUCTURE LATTICES FOR SENTENCES.
Stage 15. Construction of lattices.

The words in a sentence are listed, together with the most frequently occur- heads appertaining to them, with those words having the largest number of heads at the top. The words defined by sets of heads constitute the elements of the lattice representing the context structure of a sentence. The top point of the lattice is a latent element which is the join of all the elements in the lattice; that is, it represents all the heads contained in the lattice. The word, or words, with the greatest number of heads is the highest element in the lattice and the other words below are related to it, to the top, and to each other, by the inclusion relation which defines the nature of the lattice and which is as follows: any element ABC includes the elements AB,AC and BC, and also A,B and C. these last being the minimals. There may be a number of latent elements representing heads common to two or more elements. The bottom point of the lattice is the meet of all the elements.
(The construction of the lattices and the utilisation of the information they contain has not yet been fully incorporated into the punched card programme. In particular, there is no link between this stage and the finding of output using the semantic information on the word card.)
E.G. The lattice for GALLIA EST OMNIS DIVISA IN PARTIS TRES is given on the attached sheet.
(Ref. Masterman, Parker-Rhodes, Spärck Jones, Blackmore and Wordley "Description of the tests carried out on methods of constructing sentence lattices.)

FAN DICTIONARY.

Each output expression, (a word with all its parts of speech, or a phrase, etc.) has a range, or fan, of uses. This fan may be described in terms of a set of thesaurus heads. Every output expression is represented by a Fan Dictionary card which gives the group of thesaurus heads under which the uses may be subsumed: these heads are punched in positions 1-78 in rows 0-9.
(Ref. Spärck Jones, Blackmore and May. "A Description of the nature of Fan Cards and their position in the Punched Card Programme."

GROUP DICTIONARY.

Thesaurus heads can be grouped under a number of very general concepts which are called super heads; the head set defining an output expression

can also be represented as a group of super heads. As there are only 10 super heads many output expressions will have the same super head specification. The total set of output expressions contained in the Fan Dictionary can therefore be divided into a number of mutually exclusive groups, each characterised by a set of super heads. The groups can be separately stored and serially numbered. Each group is represented by a card with the super head set punched on it in column 80 of rows 0-9, and with the serial number of the pack stamped on it.

SUPER HEAD CARDS. H PACK.
Stage 16. First selection of output.

The word cards contain all the information necessary for a correct output; this can be obtained in two stages as follows: the super heads corresponding to the heads on the word cards are found by matching the word cards in turn with the Super Head Grid Card. On the Grid Card all the thesaurus heads are divided up into blocks, the blocks being labelled by super head numbers from 1-10. The relevant super head numbers having been found by this match, they are punched on new cards, the super head cards, in column 80 in rows 1-9.

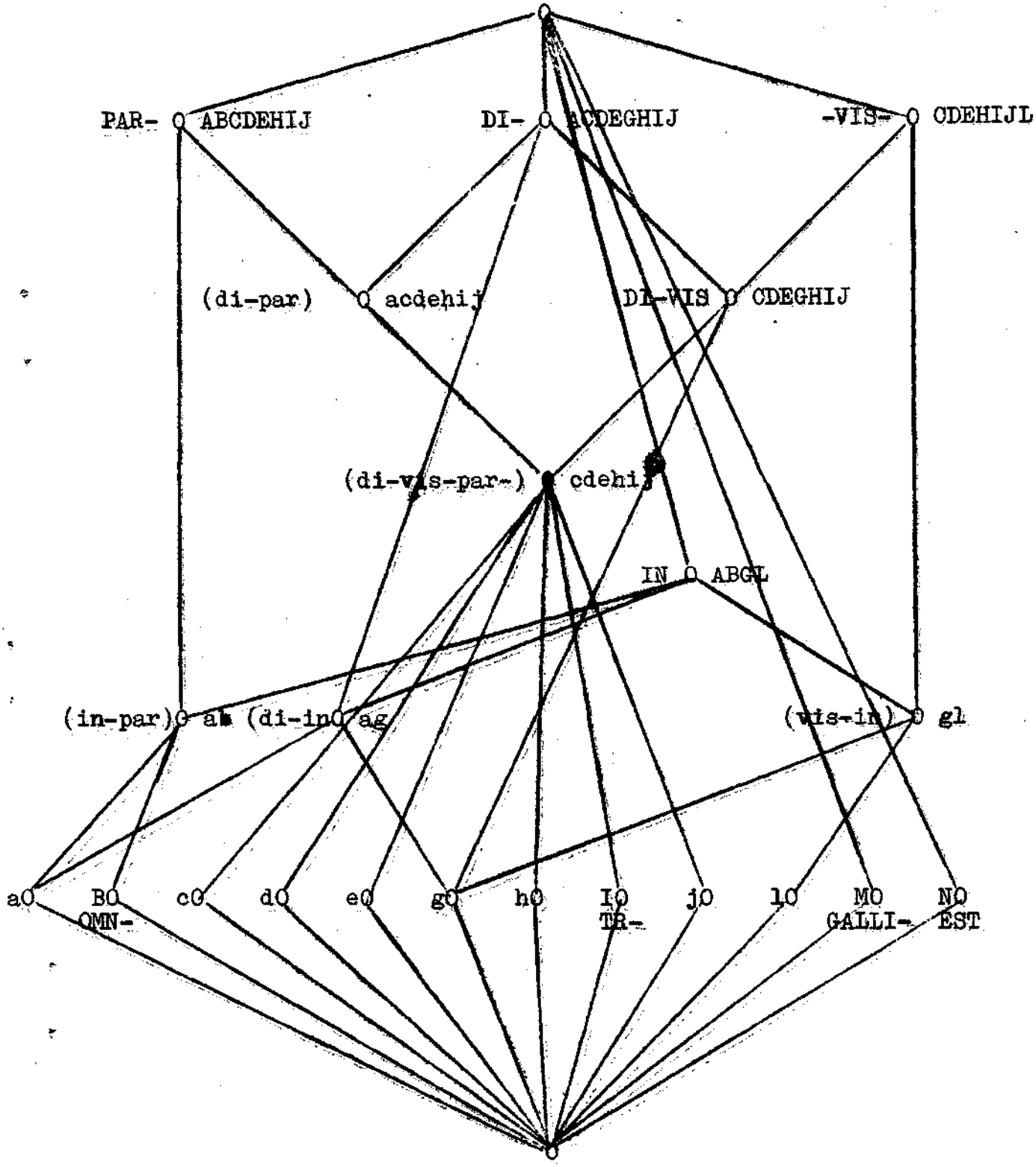
GROUP CARDS. G PACK.
Stage 17. First selection of output continued.

The Group Dictionary is sorted by the super head cards in order to obtain those Group Dictionary cards having a super head set which is the same as, or includes that generated by, the original word card for each word. The packs of Fan Dictionary cards corresponding to these selected Group Dictionary cards can be selected.

OUTPUT CARDS. O PACK.
Stage. 18. Final selection of output.
The choice of the final output expression for each word card is made by sorting the selected Fan Dictionary packs (treated as one unit) by the head combination of the word card. The card with the head set nearest, in some defined sense, that specified by the word card is the correct output. The actual word required is that written on the top of the Fan Dict. card.
Ref. Needham and Spärck Jones "A System of Two-level Access to a Translation Output using a Thesaurus.)

****

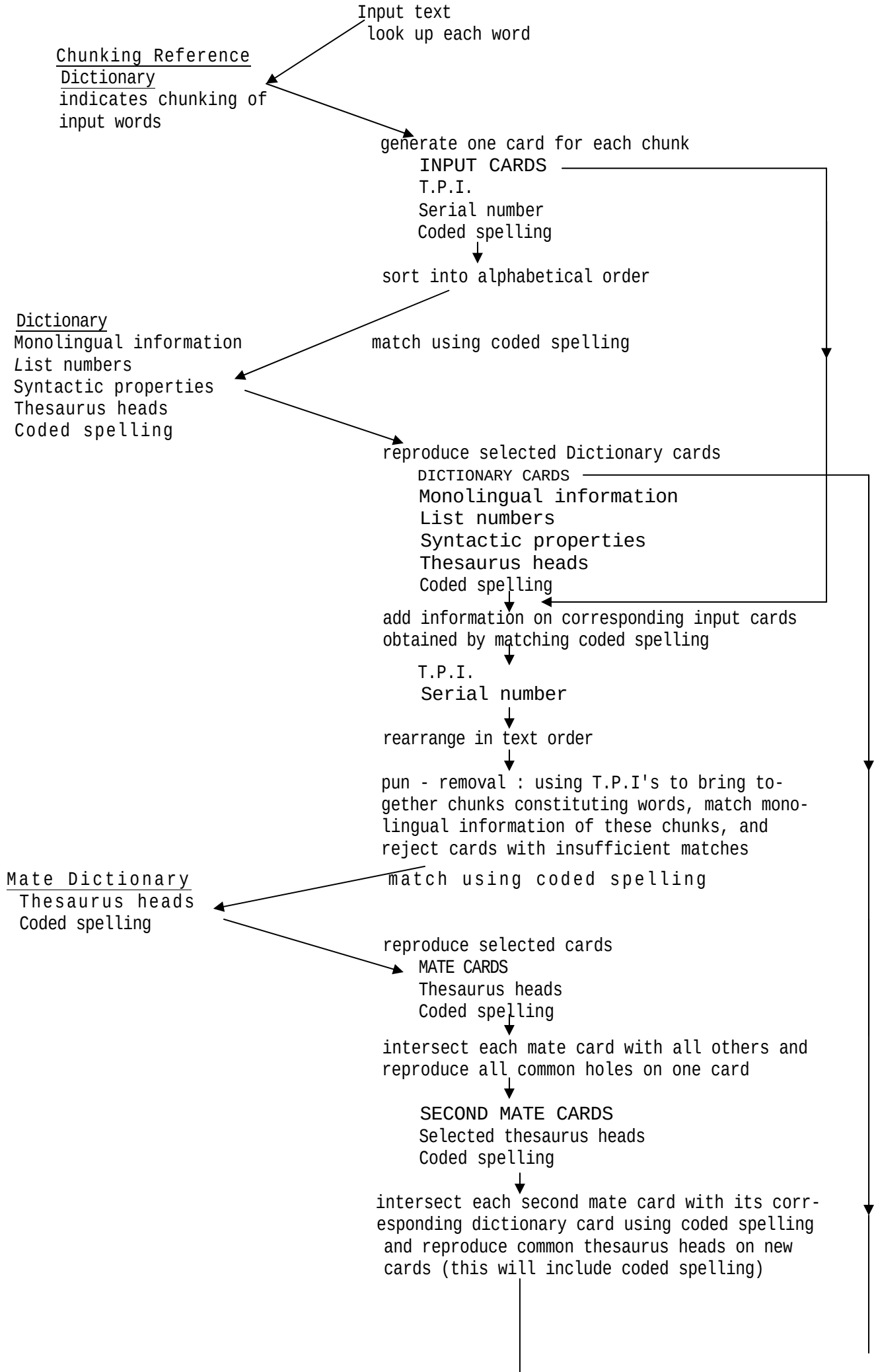
A SENTENCE STRUCTURE LATTICE, CONSTRUCTED FROM THE MOST FREQUENTLY OCCURRING SEMANTIC HEADS.

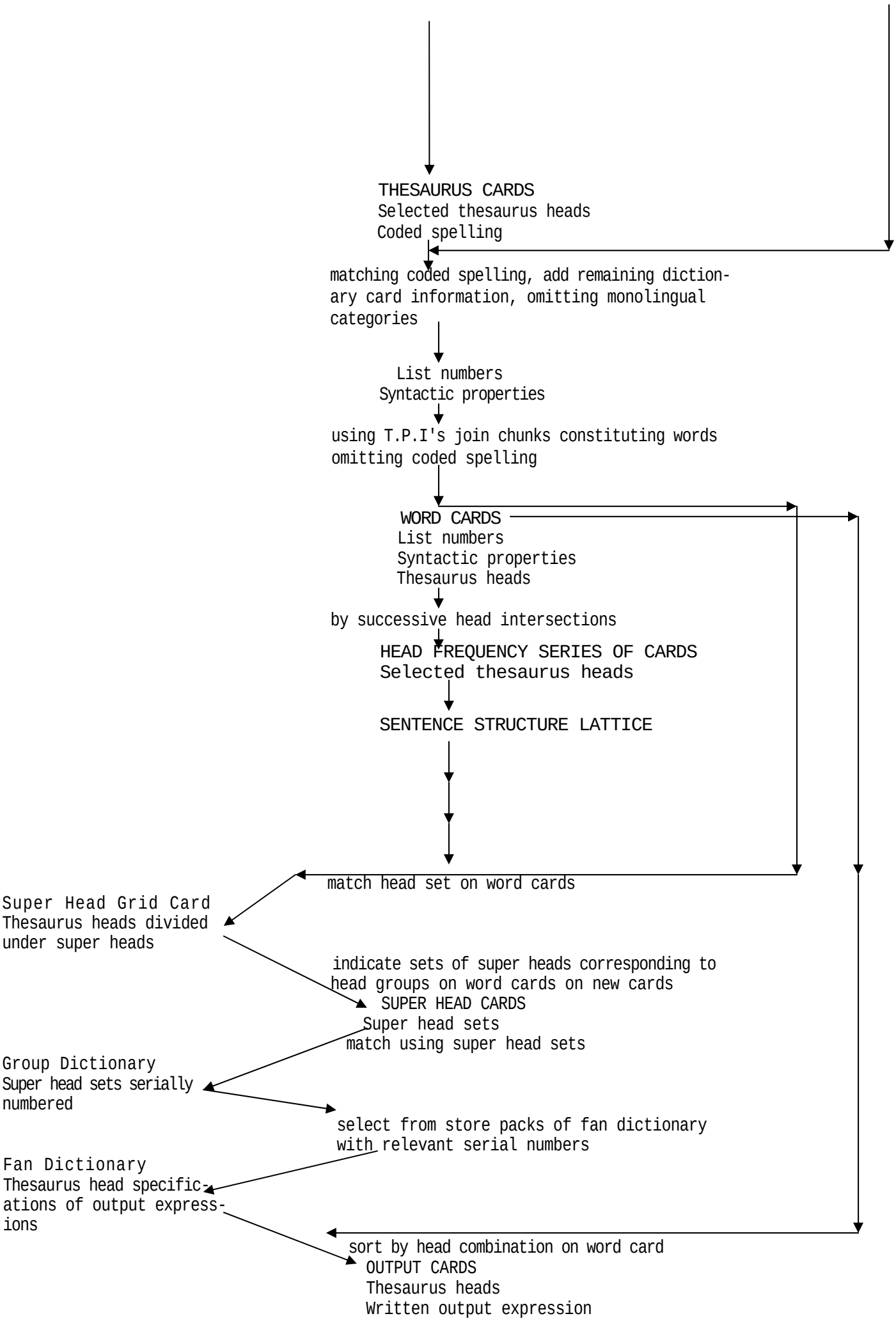


A	RELATION
B	GREATNESS.
C	DISJUNCTION
D	DECOMPOSITION
E	PART
G	ARRANGEMENT
H	DISPERSION
I	NUMBER
J	DIVERGENCE
L	APPORTIONMENT

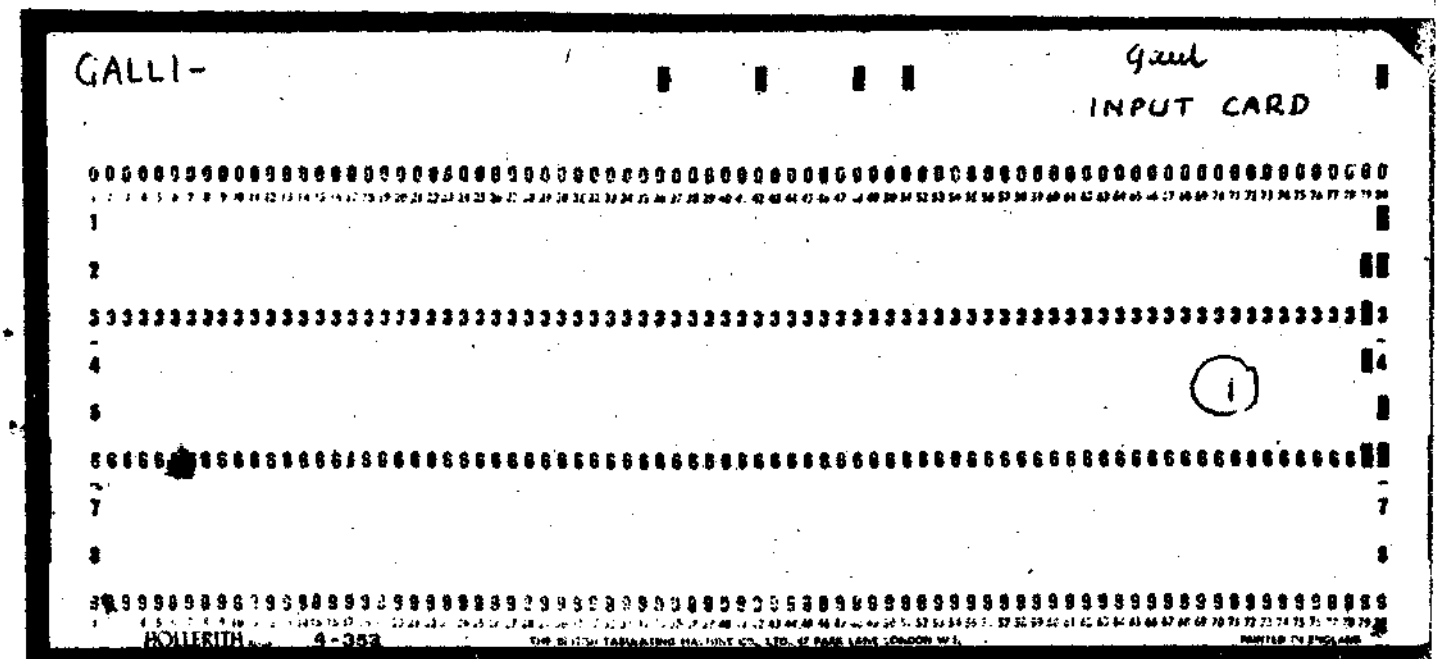
GALLI-	M
EST	N
OMN-	B
DI-	ACDEGHIJ
-VIS-	CDEGHIJL
IN	ABGL
PAR-	ABCDEHIJ
TR-	I

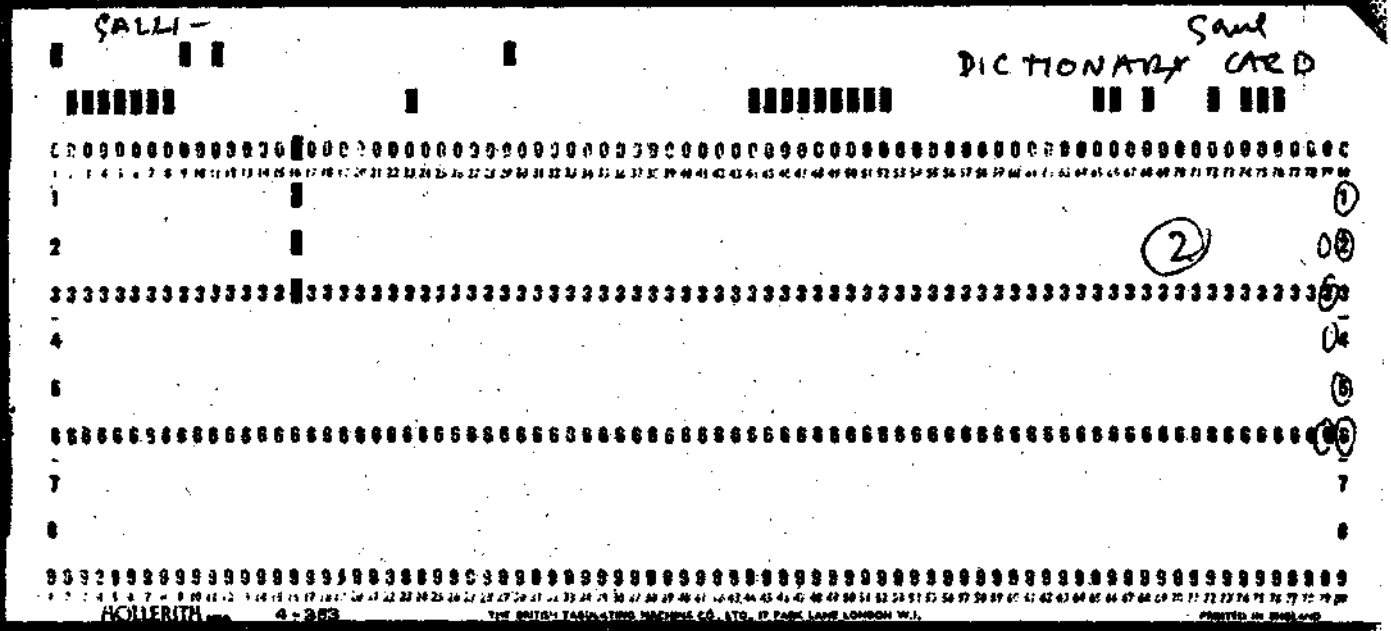
DIAGRAM OF PUNCHED CARD PROGRAMME.
Ref: K.Spärck Jones, R.M.Blackmore & C.Wordley "Library for Mechanical Translation".

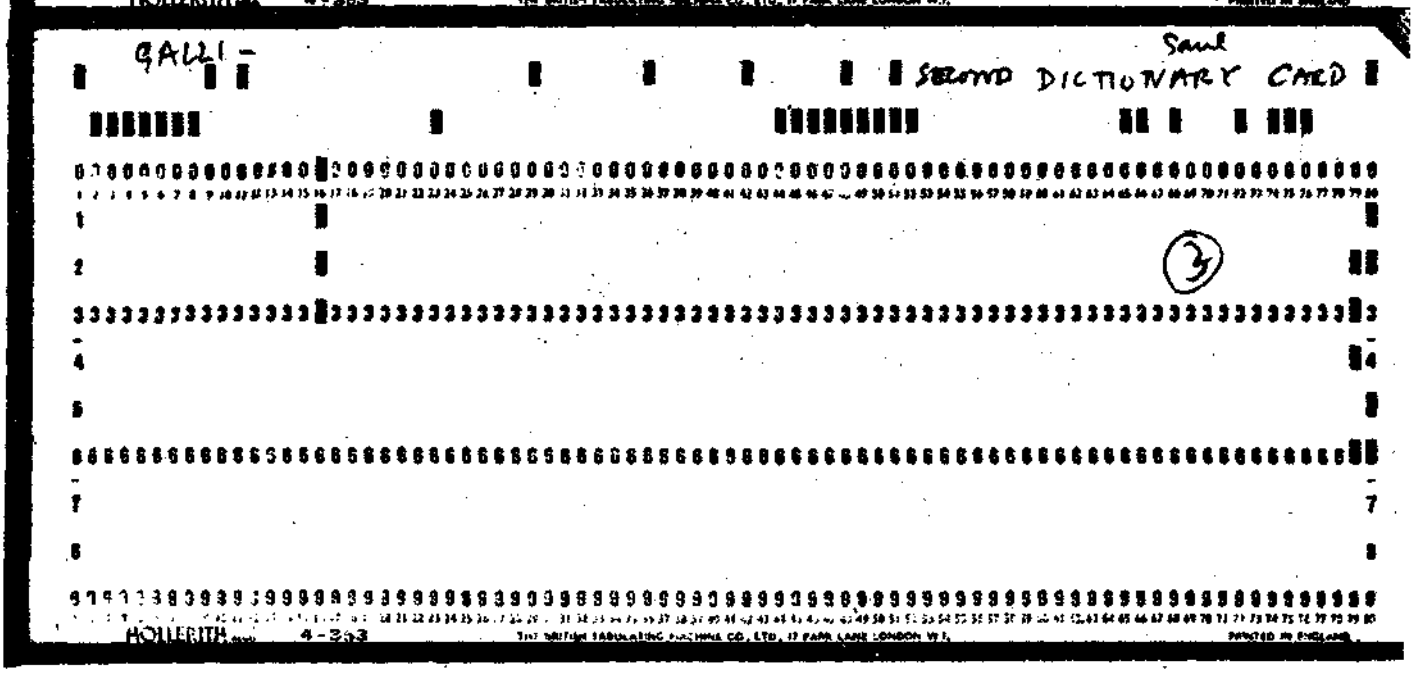


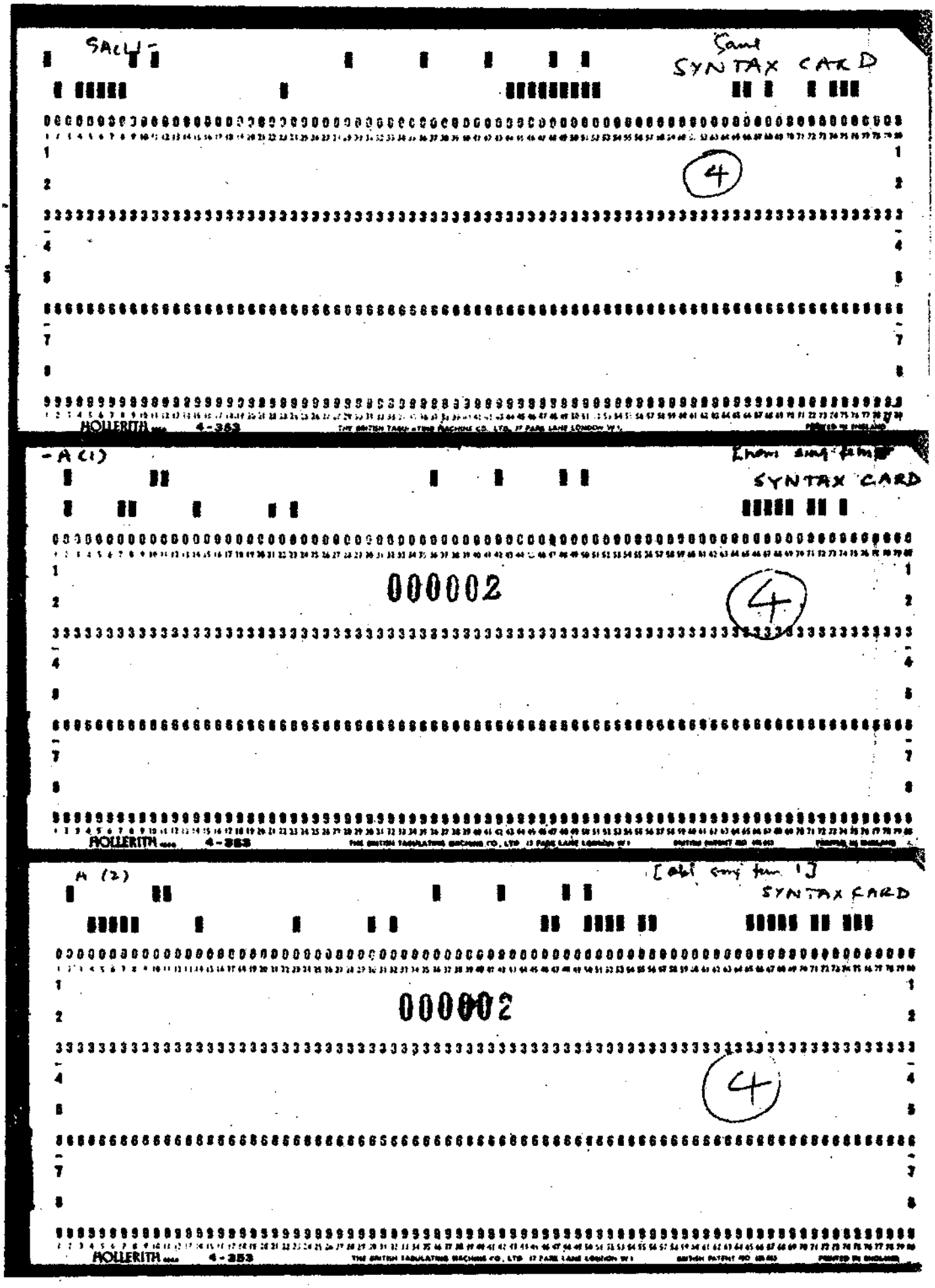

THESAURUS CARDS
Selected thesaurus heads
Coded spelling
matching coded spelling, add remaining dictionary card information, omitting monolingual categories
List numbers
Syntactic properties
using T.P.I's join chunks constituting words omitting coded spelling
WORD CARDS
List numbers
Syntactic properties
Thesaurus heads
by successive head intersections
HEAD FREQUENCY SERIES OF CARDS
Selected thesaurus heads
SENTENCE STRUCTURE LATTICE
match head set on word cards
Super Head Grid Card
Thesaurus heads divided under super heads
indicate sets of super heads corresponding to head groups on word cards on new cards
SUPER HEAD CARDS
Super head sets
match using super head sets
Group Dictionary
Super head sets serially numbered
select from store packs of fan dictionary with relevant serial numbers
Fan Dictionary
Thesaurus head specifications of output expressions
sort by head combination on word card
OUTPUT CARDS
Thesaurus heads
Written output expression

SERIES OF CARDS ILLUSTRATING PUNCHED CARD OPERATORS
AS GIVEN IN 'LIBRARY FOR MACHINE TRANSLATION'



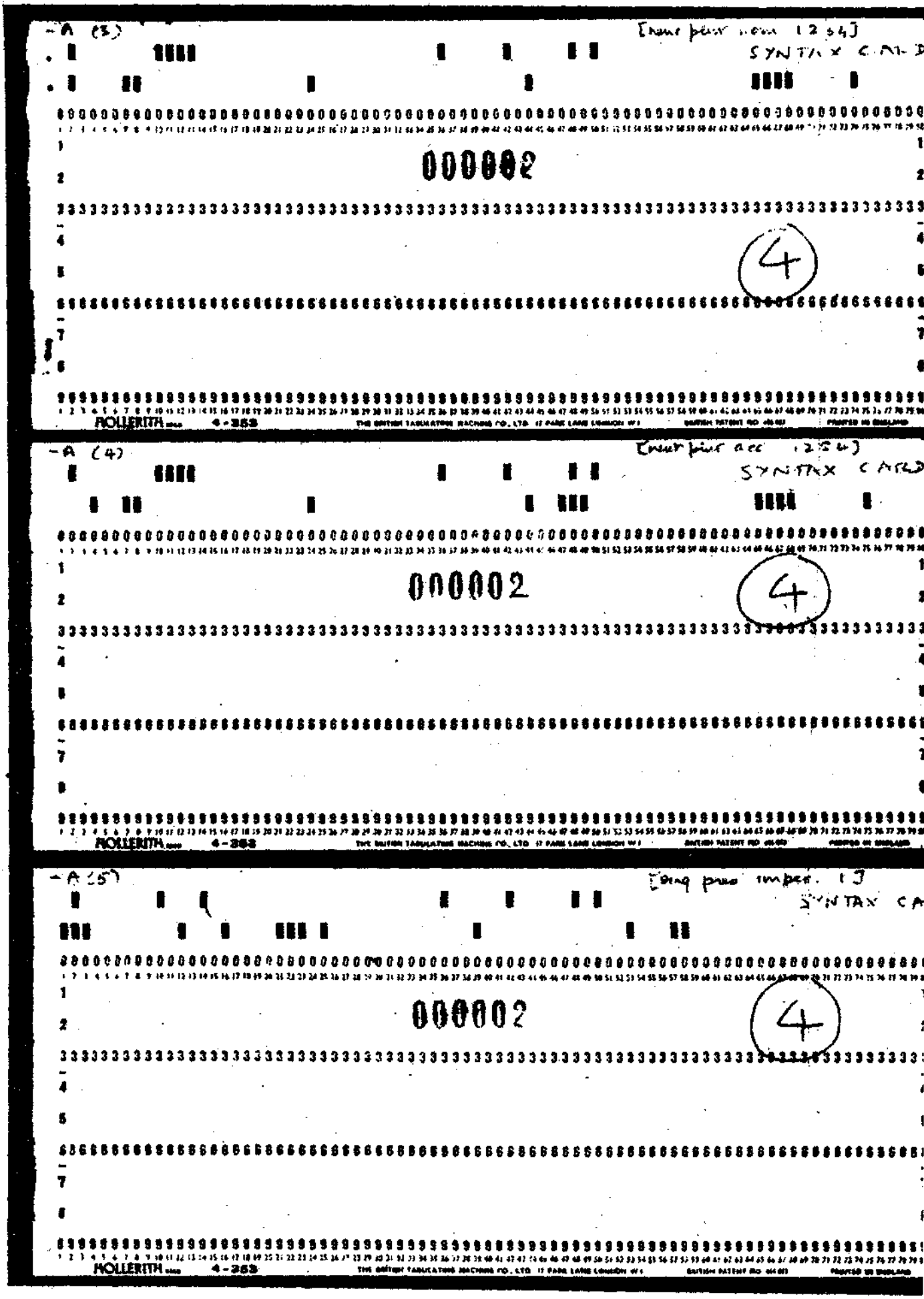





SALUT
SYNTAX CARD
4
HOLLERITH
4-363
-A (1)
SYNTAX CARD
000002
4
HOLLERITH
4-363
A (2)
SYNTAX CARD
000002
4
HOLLERITH
4-363

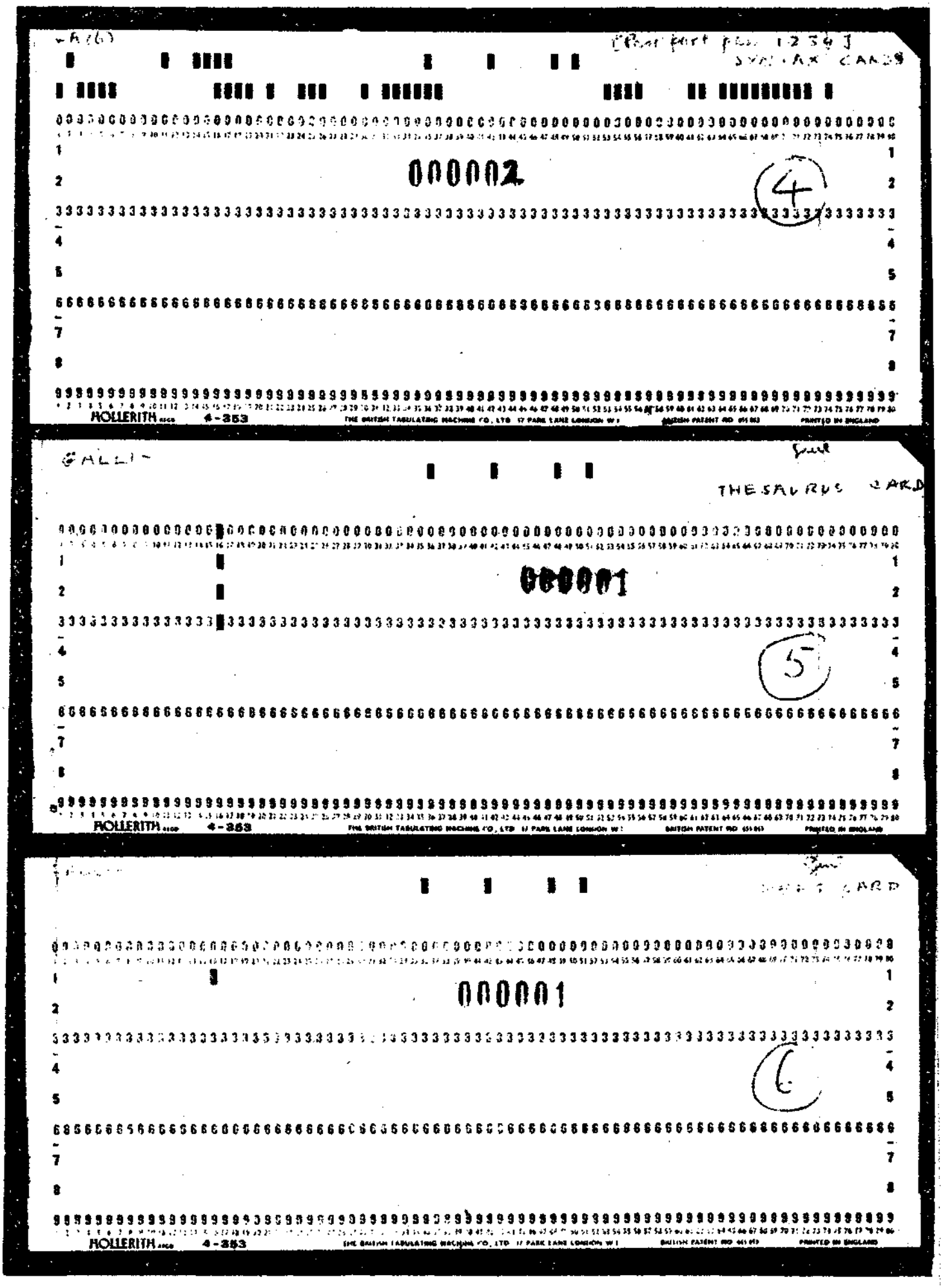
SERIES OF CARDS ILLUSTRATING PUNCHED CARD OPERATORS

AS GIVEN IN 'LIBRARY FOR MACHINE TRANSLATION'

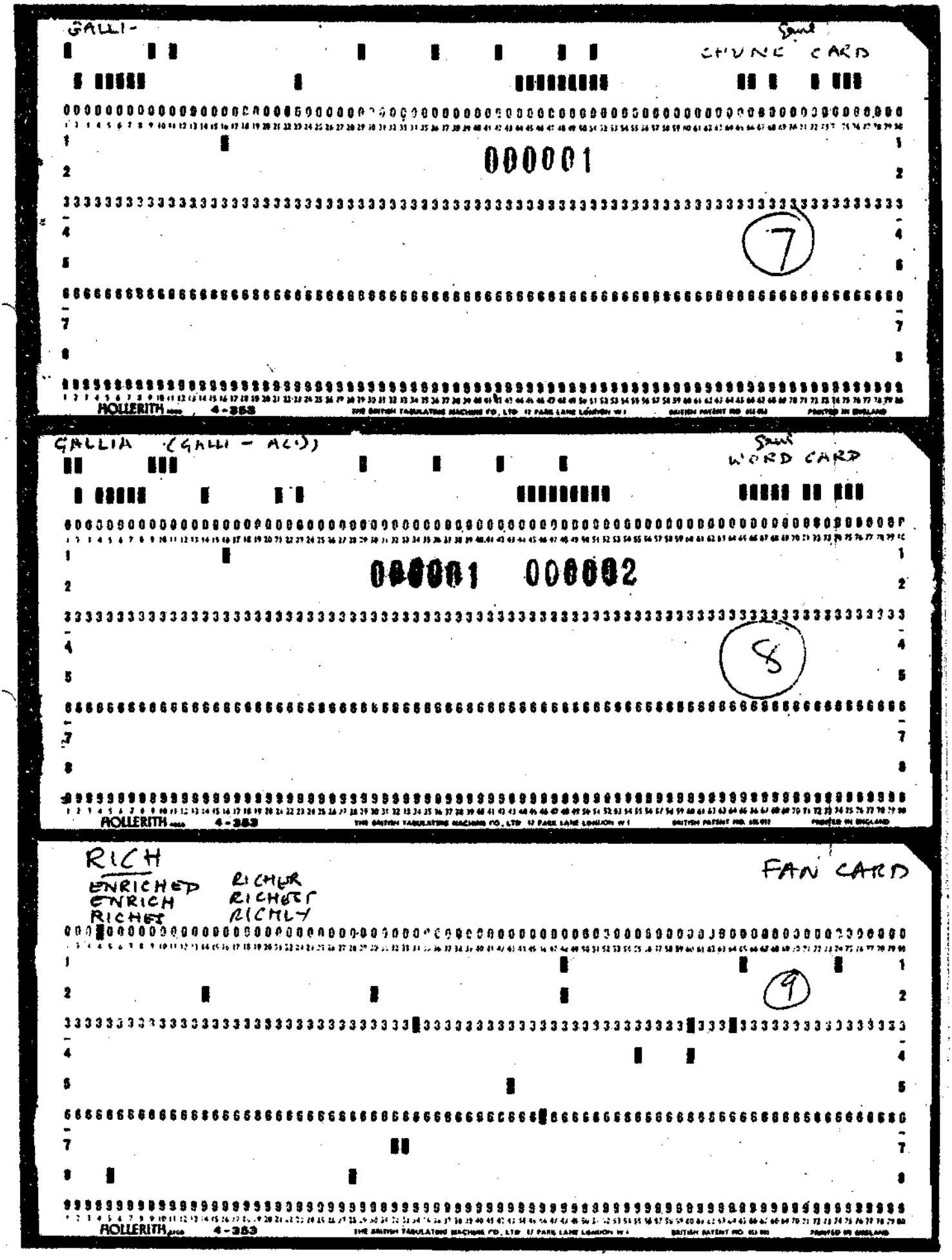


SERIES OF CARDS ILLUSTRATING PUNCHED CARD OPERATORS

AS GIVEN IN 'LIBRARY FOR MACHINE TRANSLATION'



SERIES OF CARDS ILLUSTRATING PUNCHED CARD OPERATORS
AS GIVEN IN 'LIBRARY FOR MACHINE TRANSLATION'

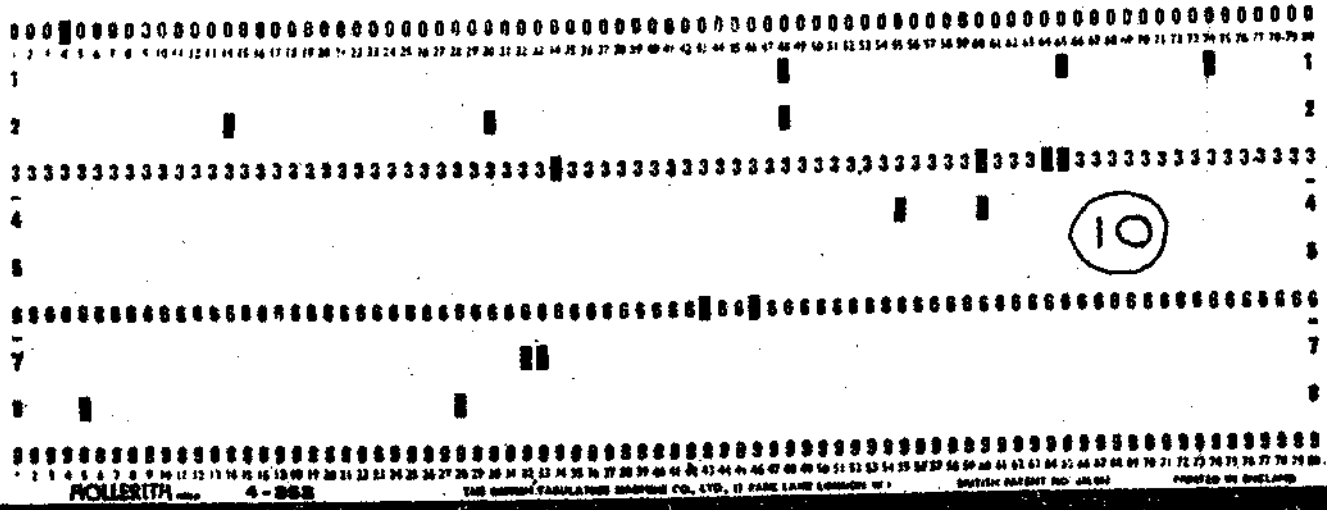


SERIES OF CARDS ILLUSTRATING PUNCHED CARD OPERATORS

AS GIVEN IN 'LIBRARY FOR MACHINE TRANSLATION'

RICH
Riches, Richer, Richest, Richly
Enrich, Enriched

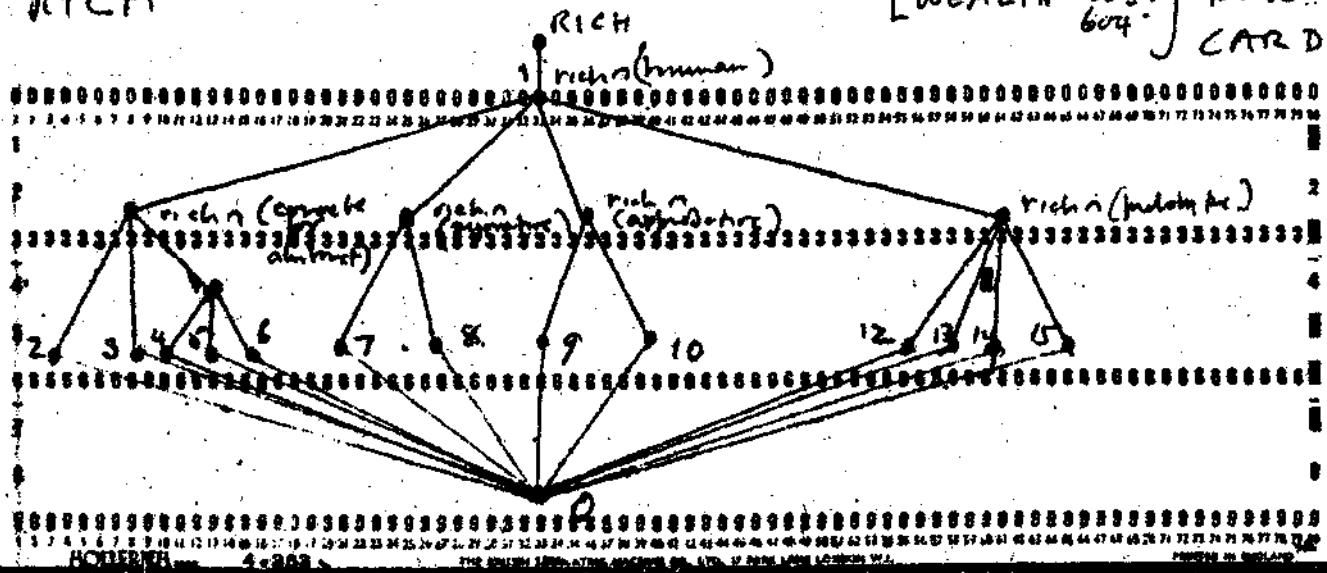
FAN OUTPUT CARD



RICH

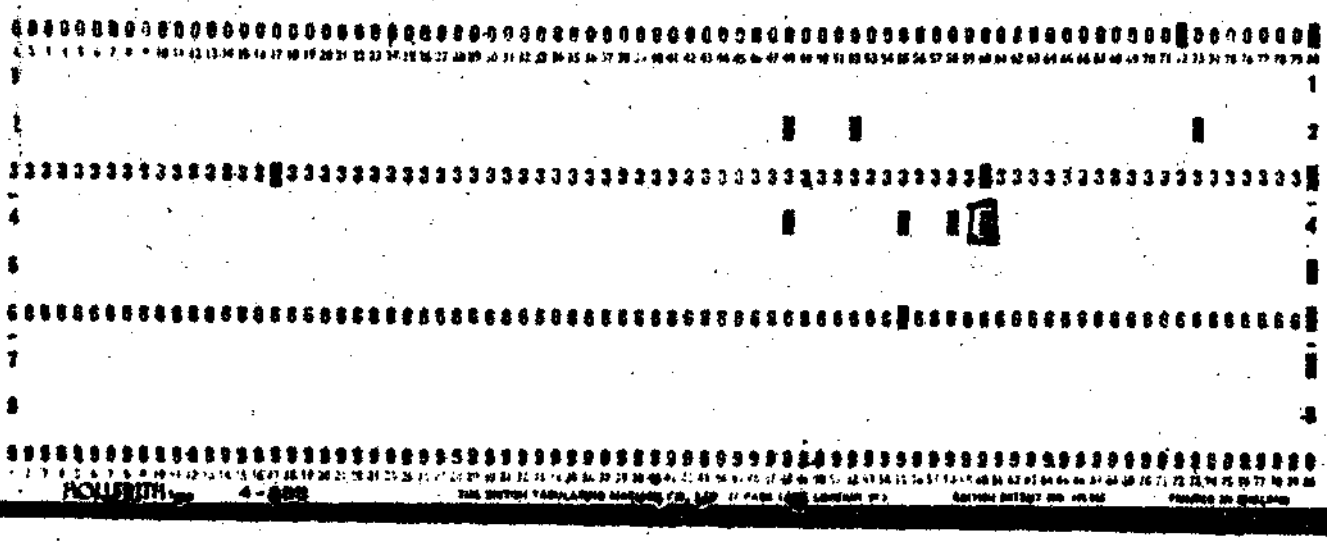
[WEALTH 803 R 604.]

FLOW CARD

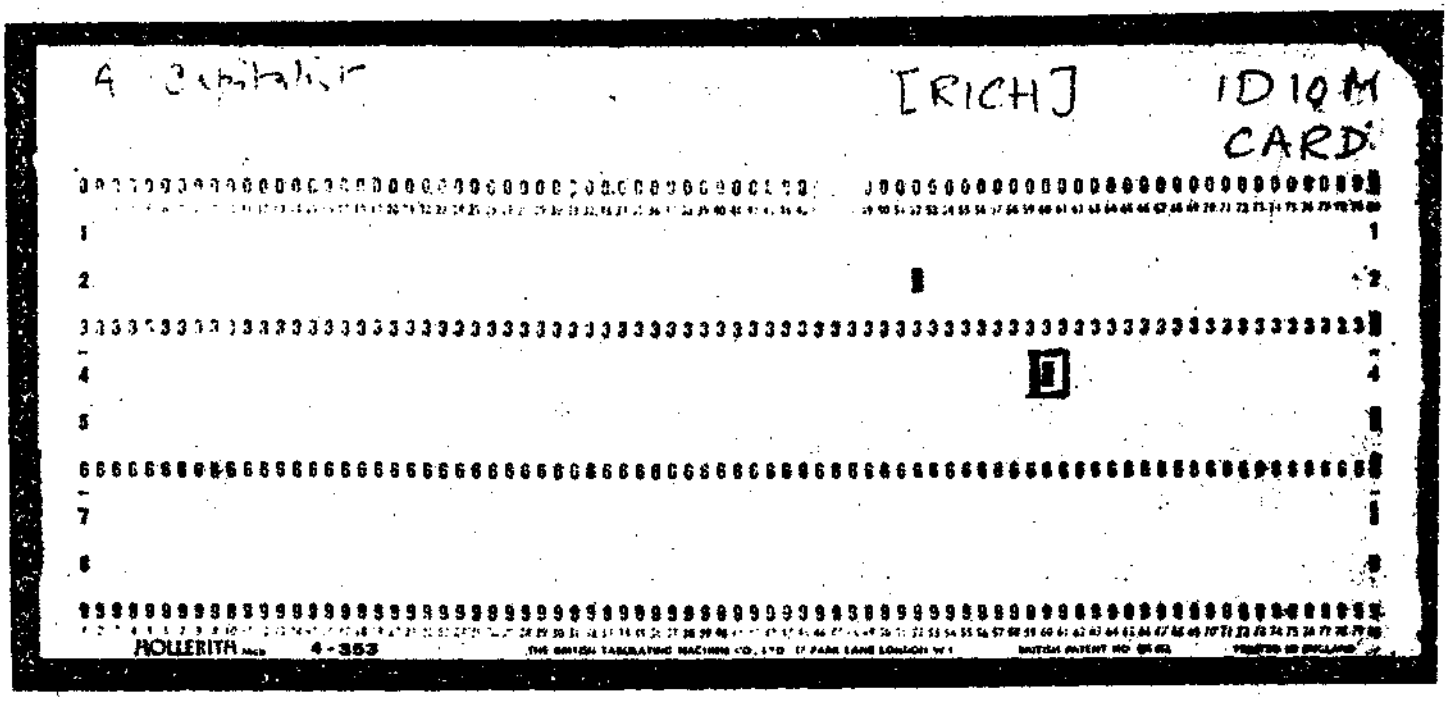


RICH

CROSS REFERENCE CARD



SERIES OF CARDS ILLUSTRATING PUNCHED **CARD OPERATORS**
AS GIVEN IN 'LIBRARY FOR MACHINE TRANSLATION'



SERIES OF CARDS ILLUSTRA.TING PUNCHED **CARD** OPERATORS
AS GIVEN IN 'LIBRARY FOR MACHINE TRANSLATION'