REPORT R-447 DECEMBER, 69

COORDINATED SCIENCE LABORATORY

PHRASE DICTIONARY CONSTRUCTION METHODS FOR THE R2 INFORMATION RETRIEVAL SYSTEM

JAMES MERRITT JANSEN, JR.

UNIVERSITY OF ILLINOIS – URBANA, ILLINOIS

PHRASE DICTIONARY CONSTRUCTION METHODS FOR THE R2 INFORMATION RETRIEVAL SYSTEM

BY

James Merritt Jansen, Jr.

This work was supported in part by the Joint Services Electronics Program (U.S. Army, U.S. Navy, and U.S. Air Force) under Contract DAAB 07-67-C-0199; and in part by OE C-1-7-071213-4557.

ACKNOWLEDGMENT

The author wishes to express his gratitude to Professor R. T. Chien for his guidance and criticism in the preparation of this thesis. Thanks also to Fred Stahl, Ph.D. candidate in the Department of Computer Science, for his help in learning the CDC 1604 computer and the ISL system.

TABLE OF CONTENTS

I. INTRODUCTION

The R2 Information Retrieval System, when complete, will be a question-answer information retrieval system in which a person, having questions over a pre-determined data base, will be able to receive answers to his questions.

This paper concerns the initial processing of the data base for the R2 system and explores initially several simple strategies for retrieving relevant statements, rather than answers, for a given question over the data base.

This paper is first concerned with the construction of a "phrase dictionary" from the data base, and, secondly, with the development of several initial attempts for information retrieval from the data base.

II. PROGRAM STRUCTURE OF R2

The R2 Information Retrieval System in the form described has been programmed on the Control Data Corporation (CDC) 1604 computer, in the Coordinated Science Laboratory of the University of Illinois. The forty-eight programs and subprograms comprising the R2 Information Retrieval System were written in the Information Search Language system (ISL) and in the Illinois Assembly Resident system (ILLAR). Use was also made of the CDC CO-OP Monitor System for a generalized alphabetic sort program. Of the forty-eight routines written, thirty-eight are used in the construction of the "phrase dictionary." Several of these routines are for the purpose of supplying status information and action requests to the computer operator. The remaining ten programs and subroutines are incorporated in the strategies for retrieving relevant statements in response to questions asked over information in the data base of the R2 system. Appendix A contains brief descriptions of the computer programs, flow charts, and operating instructions for the computer programs for the construction of the "phrase dictionary". Appendix B contains operating instructions for the use of the initial strategies of the R2 system.

III. COMPOSITION OF THE INITIAL R2 RETRIEVAL SYSTEM

The initial R2 Information Retrieval System is composed of (1) a data base, that is, a set of sentences, paragraphs, and statements over some subject matter; (2) a "phrase dictionary" constructed from the edited data base; (3) a set of strategies for retrieving relevant statements from a set of questions; and (4) a user seeking answers to questions over information contained in the data base.

The data base used in the experimental work performed on the R2 system is described in chapter IV. The construction of the "phrase dictionary" is described in chapter VI. An example of a portion of the dictionary is given in Appendix C. The six basic strategies composing the initial strategy set are discussed in chapter VII.

IV. CHOICE OF A DATA BASE

For the study of the information retrieval method described herein, the official Illinois driver's manual "Rules of the Road" was chosen for the data base. From this the descriptive "R2" was taken to describe the information retrieval system discussed. The choice of "Rules of the Road" for the data base was twofold: (1) this data base contains factual information and (2) this data base is of such length as to completely test the generality of the R2 information retrieval system.

"Rules of the Road" consists of 94 pages of factual information--information of the type a person interested in obtaining a driver's license in Illinois would need and want to know. A person seeking answers to questions concerning traffic rules and regulations in Illinois could definitely profit from the R2 retrieval system applied to "Rules of the Road" as its data base.

V. EDITING OF THE ORIGINAL DATA BASE

The original data base requires editing so as to conform to certain restrictions imposed by the Information Search Language (ISL) and also to handle other problems described below. The ISL system makes use of the following symbols: the minus sign, hyphen, and dash (-); the semicolon (;); and the colon (:). Hence, the original data base containing such symbols must be edited to remove these symbols. The first step in the editing process is to remove all occurrences of dashes, hyphens, and minuses. The symbols are either replaced by commas or spaces depending on the context in which they appear. The semicolon is replaced by a comma or the phrases separated by the semicolon are broken into two sentences. Here again, the option depends on the context in which the symbol is used. Quotation marks are completely removed. Question marks in the data base are coded with the symbol "+".

The use of the colon and the use of ellipsis periods to indicate lists occurs frequently in the "Rules of the Road." For example, consider the statement from the "Rules of the Road" shown in Figure 1 at the top of the next page.

The coded form of the statement shown in Figure 1 is shown in Figure 2. Note that abbreviations, quotation marks, and hyphens have been removed conforming to the rules of editing outlined above.

PASSING IS PROHIBITED ON A TWO-LANE HIGHWAY...

In a no-passing area marked by a yellow strip or a "DO NOT PASS" sign.

On a hill or a curve where a vehicle approaching from the opposite direction might create a hazard.

Within 100 ft. of an intersection or railroad crossing.

Figure 1

Example for the illustration of the coding of list structures.

PASSING IS PROHIBITED ON A TWO LANE HIGHWAY IN A NO PASSING AREA MARKED BY A YELLOW STRIPE OR DO NOT PASS SIGN.

PASSING IS PROHIBITED ON A TWO LANE HIGHWAY ON A HILL OR A CURVE WHERE A VEHICLE APPROACHING FROM THE OPPOSITE DIRECTION MIGHT CREATE A HAZARD.

PASSING IS PROHIBITED ON A TWO LANE HIGHWAY WITHIN 100 FEET OF AN INTERSECTION OR RAILROAD CROSSING.

Figure 2

Edited form of Figure 1.

The purpose of coding lists in this format should become quite obvious when the construction of the dictionary is explained below.

The above example does not demonstrate the removal of commas or decimal points from numbers. The comma and the period are used in defining normal breaks in the text in the edited sentence structure. This should explain the need for the removal of these symbols in numbers.

Following the procedure outlined above, the edited text of the data base, in this case "Rules of the Road," is punched on cards. The card images are then transformed into card image magnetic tape records for the "phrase dictionary" construction phase of the R2 system.

VI. CONSTRUCTION OF THE "PHRASE DICTIONARY"

The "phrase dictionary" is very different from an ordinary dictionary containing usages, definitions, synonyms, etc. The "phrase dictionary" consists of groups of words (termed "phrases") with a list of numbers which reference sentences in which the phrases occur. The phrases follow a set of rules described below.

The card image records stored on magnetic tape are converted to sentence image records and numbered sequentially in ascending order of their appearance in "Rules of the Road". A WIS index is performed on the sequentially numbered sentences of the data base. The sequentially numbered sentences now constitute the new data base which will hereafter be referred to as the data base. WIS means "word in sentence". The name is derived from the symbol "KWIC" (Keyword in Context) where here <u>all</u> words, not just keywords, are used in the construction of the index and the context is a <u>sentence</u>. This portion of the construction phase is performed in "Phrase 1" of the program "Phrase" and the sort following.

Consecutive entries of the WIS index are compared to find the longest string of words between these entries which match. For example, consider the following WIS index entries shown in Figure 3 at the top of the next page.

Shown in Figure 4, are the entries corresponding to the WIS index information shown in Figure 3 which have been processed to find the maximal match between consecutive entries. The programming for

this step of the construction of the dictionary is written in "Phrase 2" of the program "Phrase."

...STOP AT A FLASHING RED LIGHT...	437
...STOP AT A RED LIGHT...	239
...STOP AT A STOP SIGN...	625
...STOP BEFORE BACKING INTO THE STREET...	22
...STOP BEFORE CROSSING THE TRACKS...	1007
...STOP BEYOND...	761

Figure 3

A portion of the WIS index.

STOP AT A 437,239,625

STOP BEFORE 22,1007

STOP 761,1007

Figure 4

"Phrase 2" application to WIS index segment shown in Figure 3.

The numbers on the WIS index entries in Figure 3 and the partially complete phrase dictionary entries shown in Figure 4 represent the sentences in which the respective entries appear.

The removal of "prefixes" from the previous stage is the programming effort involved in "Phrase 3" of "Phrase." A "prefix" is defined in the following manner. If an entry is the beginning of another entry, and if these two entries have any sentences numbers in common, then the first entry, since the entries are in alphabetic order, is considered a "prefix." The numbers in common between entries are removed from the prefix entry. If the prefix entry results with all sentence numbers removed, that entry is deleted since it is entirely contained in a larger phrase. For an example of the programming effort involved in "Phrase 3" consider the partially complete dictionary entries shown below in Figure 5.

A LEFT	571,587,718,1081
A LEFT TURN	571,587,624,657,1081
A LEFT TURN INTO	587,657
A LEGAL RESIDENT OF	24,240,243
A LEGAL RESIDENT OF ANOTHER STATE	240,243

Figure 5

A portion of the results from "Phrase 2."

The output of "Phrase 3" operating on the input shown in Figure 5 appears in Figure 6 at the top of the next page.

A LEFT	718
A LEFT TURN	571,624,1081
A LEFT TURN INTO	587,657
A LEGAL RESIDENT OF	24
A LEGAL RESIDENT OF ANOTHER STATE	240,243

Figure 6

"Phrase 3" application to the data of Figure 5.

Removal of "suffixes" from the output of the "Phrase 3" constitutes the programming segment "Phrase 4." An entry is said to be a "suffix" if the entry is the ending of another entry and if the two entries have sentence numbers in common. To facilitate the removal of suffixes in the phrase dictionary which has been constructed thus far, each entry is spelled backwards and the entries are sorted alphabetically in "Phrase 3". "Phrase 4" then operates on this sorted structure to remove the suffixes and restore the resultant dictionary entries to normal spelling. An example of the input to Phrase 4 with the resultant output is shown in Figures 7 and 8.

EB	34,35,36,36,37,41
EB LLIW	34,36,37,41
EB LLIW EREHT	34,36
EB LLIW EREHT DNA	36
EB LLIW SREVIRD	41
EB TSUM	35,38,45
EE TSUM NOITCHRTSNI	35,38

Figure 7

Partial Input Segment to Phrase 4.

WILL BE	37,41
THERE WILL BE	34
AND THERE WILL BE	36
DRIVERS WILL BE	41
MUST BE	45
INSTRUCTION MUST BE	35,38

Figure 8

Phrase 4 output of the
partial input list of Figure 7.

Following an alphabetic sort to restore the phrases obtained by the removal of suffixes, the sentence numbers are sorted and the format of the records are changed to give the completed phrase dictionary. Of the 1333 sentences entries in the data base, 5855 dictionary entries were found. Each entry is maximal in length since suffixes and prefixes are removed. The dictionary and data base are then ready for use by the strategies for retrieving relevant statements in answer to questions over the textual information contained in "Rules of the Road."

VII. THE INITIAL STRATEGIES OF R2

The three basic initial strategy techniques which have been developed, programmed, and tested are appropriately labelled strategy 1, strategy 2, and strategy 3, in the order of their development. Strategy 1 is a base from which strategy 2 and strategy 3 are developed.

Strategy 1, when given a statement or question, first sets up a list of pointers which point to the first character of every word in the question. For example, consider the question in Figure 9 below which is used to describe strategy 1.

```
1   2   3   4     5  6   7
↓   ↓   ↓   ↓     ↓  ↓   ↓
WHAT ARE THE RULES OF THE ROAD+
```

Figure 9

Question with associated word pointers.

The pointers are then arranged in the order of the alphabetically sorted words. Hence, the order of the pointers for the example edited in Figure 9 will be 2, 5, 7, 4, 6, 3, 1. Note that not only are single words used to determine the ordering of the pointers, but also the words following. The purpose of arranging the pointers in the alphabetic sort order of the words in the text is to perform the dictionary lookup with a single pass of the magnetic tape on which the alphabetically sorted dictionary entries are stored. Consider pointer 2 which points to the letter "A" in the word phrase "ARE THE RULES OF THE ROAD+". The dictionary is searched for the entry "ARE." If "ARE" is found in the

dictionary, the sentence numbers associated with the entry "ARE" are saved. The dictionary is searched for "ARE THE," if "ARE THE" is found, the sentence numbers associated with "ARE" are replaced by those associated with "ARE THE." This process is continued until a maximal phrase dictionary entry is found for the associated pointer. The sentences numbers are retained only for the maximal phrase entry obtained for the particular pointer. This process of finding maximal phrases associated with each pointer is continued until all pointers and their associated work phrases have been retrieved. The sentence numbers retrieved for all pointers are sorted numerically and a one-pass retrieval of the sentences is made from the data base. Although strategy 1 is very poor, it is much better than a straight word lookup technique. This is the postulate upon which the phrase dictionary construction technique was formulated.

Using the results of strategy 1, strategy 2 retains only those maximal phrases and the associated sentence numbers which are not subphrases of a larger phrase. For example, consider again the question cited in Figure 9 and assume that "RULES OF THE ROAD," "OF THE ROAD," "THE ROAD," and "ROAD" are the maximal phrases associated with pointers 4, 5, 6, and 7 respectively. Here "ROAD" is a subphrase of "THE ROAD," "THE ROAD" is a subphrase of "OF THE ROAD," and "OF THE ROAD" is a subphrase of "RULES OF THE ROAD." By strategy 2, only the sentence numbers associated with "RULES OF THE ROAD" would be retained for pointers 4, 5, 6, and 7. The sentences associated with the remaining sentence numbers are retrieved as the statements relevant to the question asked.

Strategy 3 retains only those sentence numbers associated with the largest (most characters) maximal phrase. The sentences retrieved are associated with a single phrase as opposed to several maximal phrases with the other strategies.

Another strategy option (threshold of 2) which increases the total number of strategies to six is the following. Using one of the three strategies above, only retrieve those statements which are referenced two or more times. This deletes many of the would-be retrieved statements, but also often deletes a relevant statement.

A second strategy option (ENS) which increases the total number of possible strategies to twelve is the following. Note that this option may be used with the option above. In the case in which both options are exercised, the second strategy option is performed first and the first option second. The second strategy option retrieves sentences numbered n-1, n, n+1 where n represents a sentence number retrieved by strategy 1, strategy 2, or strategy 3, which ever is applied to the question. This option retrieves more information, while the first reduces the amount retrieved. The idea here is that, in the text, sentences which lie adjacent to a given sentence will possibly contain additional information relevant to the given sentence because of the paragraph structure of the text.

A strategy relevancy rate to determine which of the twelve strategies is the best strategy has been defined. This relevancy rate (R_1) is

$$R_1 = \frac{N}{E+N},$$

where

R_1 = relevancy rate,

N = number of relevant statements retrieved, and

E = number of irrelevant statements retrieved.

Using this strategy relevancy rate, strategy 2 with the option Threshold of 2 is the best strategy. This strategy has a relevancy rate of approximately .25 for some 20 questions asked over the data base. However, the best results for a particular question may come from use of a different strategy than strategy 2 with a threshold of 2.

Although the relevancy rate is quite low using the present technique, a second relevancy rate has been defined. The purpose of the addition of a second relevancy rate is that R_1 does not penalize a particular strategy for retrieving only a few of the relevant answers contained in the data base. The second strategy relevancy rate R_2 is defined to penalize a strategy which does not retrieve <u>all</u> relevant answers to the question which is contained in the data base. The formula for R_2 is

$$R_2 = \frac{N}{A+E},$$

where

R_2 = strategy relevancy rate,

N = number of relevant statements retrieved,

E = number of irrelevant statements retrieved, and

A = number of relevant statements to the question contained in the data base.

The relevancy rates are computed by hand for R_2. This is a major effort since the entire data base must be scanned for all statements relevant to a particular question.

A number of examples of using the strategies discussed are given in Appendix D.

VIII. CONCLUSION AND PROPOSAL FOR RESEARCH

The work presented in this paper represents the first step in the development of the R2 question-answering information retrieval system. It indicates that information retrieval systems with natural language communication on all levels can be successfully implemented. The initial R2 system has already introduced a number of new features that will aid in the implementation of the total system.

First, the system is data-base independent. That is, the implementation described would have worked equally well on any other coherent textual data-base without any modification to the programs. Strategy 2 with a thr old of 2 has $R_1 = .25$.

Second, the maximal phrase concept has defined a new approach to natural language interpretation. The questions posed to the system were not analyzed on a syntactic level. This means there was no attempt made to determine the structure of the question other than the relationship of the constituent maximal phrases to the statements of the data-base. This technique has the obvious advantage of not acquiring the ambiguity involved in syntactic analysis and yet extracting important relationships between the question and the data-base.

There are a number of ways in which the present system may be clearly improved: (1) Allowing paragraphs and chapters to be treated as unifying entities, (2) Recognition of hyphenated words, (3) Replacement capability to accommodate synonyms with possible relational structures utilizing feedback, and (4) Sensitivity to pluralization and other variant word forms.

The next phase of development should deal with methods of bringing into play syntactic analysis in the various levels of the system; for interpretation of the question, for structuring the data base, and for the eventual process of inducing and synthesizing the resultant responses of the system. The addition of this capacity should greatly enhance the ability of the system to deal with natural languages.

REFERENCES

Lee, Donald, ILLAR, Coordinated Science Laboratory, Urbana, Illinois, 1968.

Kelley, Karl, Ray, Sylvian R., and Stahl, Fred, Information Search Language, Department of Computer Science, File No. 735, Urbana, Illinois, September 12, 1967.

Powell, Paul, "Rules of the Road," Springfield, Illinois, 1968.

SORT/REFERENCE Manual, Control Data Corporation, Palo Alto, California, January, 1964.

APPENDIX A

DOCUMENTATION OF THE SYSTEM

Appendix A contains the following information in the order listed below:

(1) a brief description of the programs for the construction of the "phrase dictionary",

(2) flow charts of the programs for the construction of the "phrase dictionary", and

(3) instructions for the use of the construction programs for the CDC 1604 computer in the Coordinated Science Laboratory.

MAIN PROGRAMS

PHRASE1 - This main program calls the following programs for execution in the order given:

1. MESG1
2. DELPG
3. MAKES
4. THREEPA
5. NUMBER
6. FXRLG1
7. QADPRE
8. PREPSORT
9. MESE2

PHRASE1 is used to edit the input and made sentences of the edited input data. Following the editing, the WIS Index is produced from the sentences.

PHRASE2 - This main program which follows the first alphabetic sort following PHRASE1, calls the following programs:

1. MESG1
2. PHRASER
3. COMBINE
4. REMTAIL
5. REMPREF
6. FXRLG2
7 PREPSORT
8. MESG2

PHRASE2 first determines all maximal phrases in the data base. Prefixes are then removed.

PHRASE3 - This main program which follows a CO-OP monitor sort removes duplicate prefix entries from the output of PHRASE2 and sets up the phrases in backwards form for the removal of suffixes. The program subroutines incorporated in PHRASE3 are as follows:

1. MESG1
2. REMDUP
3. BACKADD
4. FXRLG2
5. PREPSORT
6. MESG2

PHRASE4 - This main program which follows the third sort is made up of the following subroutines:

1. MESG1
2. REARNGE
3. REMSUFF
4. REMDROP
5. MESG2

PHRASE4 removes the suffixes from the data base and prepares the output records in final form.

PHRASE5 - Following a special sort at the end of PHRASE4, the phrase dictionary is in completed final form. PHRASE5 puts the output of NUMBER in final form. PHRASE5 consists of the program RENUMBER.

SUBROUTINES

DELPG - This subroutine removes the page number from each ISL card image record. The input records are the output records of IBMISL1.

MAKES - This subroutine follows DELPG and removes all occurrences of multiple blanks as it constructs sentence image records. All sentences will end with a period(.), a question mark(coded as +), or ellipsis periods(...).

THREEPA - This subroutine operates on the output of MAKES to append all phrases beginning with an asterisk to the previous sentence having ellipsis periods. The ellipsis periods are removed with the asterisks, and a blank is inserted between phrases. All occurrences of minuses, dashes, and hyphens (-) are removed and replaced with blanks.

NUMBER - The subroutine numbers each record sequentially by placing a 48 bit BCD number at the beginning of each sentence.

FXRLG1 - The subroutine operates on each record of the output of NUMBER to produce records of a fixed length. The fixed length is equal to the length of the longest record from the output of NUMBER. The records are left justified with blanks right to fill out the record.

QADPRE - Operating on the output of FXRLG1, this subroutine produces the WIS index. The output records are left fixed in length with a length equal to the length of those in FXRLG1.

PREPSORT - This subroutine sets up the parameters and calls SORTPROG to write a CO-OP Monitor file to use the CO-OP Sort Routines.

SORTPROG - This subroutine is a modified ILLAR subroutine to write a CO-OP Monitor record to be used for the CDC CO-OP Monitor sort routines.

PHRASER - This subroutine uses the fixed length alphabetically sorted sentence records from PHRASE1 as input. Consecutive sentences are compared to find the longest character match. When a match is found, the phrase and the associated sentence numbers are written. Each entry is also compared to the previous entry written. If it is identical, the sentence number alone is written.

COMBINE - This subroutine takes the output of PHRASER and appends to the phrase entry, with its two sentence numbers, all those sentence numbers which follow up to the next phrase.

REMTAIL - This subroutine removes the blanks from the end of the phrases, shortening the record length, and moves the sentence numbers immediately following the phrase.

REMPREF - This subroutine operates on the shortened REMTAIL output to remove prefixes.

FXRLG2 - This subroutine fixes the record length in preparation for a CO-OP Sort. The phrases are fixed in length as well as the sentence numbers.

REMDUP - This subroutine operates on the sorted output of PHRASE2 to remove duplicate prefix entries.

BACKADD - This subroutine follows REMDUP and inverts the prefix phrasers character for character. The output records contain both the backward and forward phrases and their sentence numbers.

REARNGE - This subroutine rearranges the sorted output of PHRASE3 to put the phrases in the final form. The BCD sentence numbers are converted to binary.

REMSUFF - This subroutine removes the suffixes from the data base by removing sentence numbers.

REMDROP - This subroutine drops the backwards phrase from the output of REMSUFF.

RENUMBER - This subroutine operates on the output of NUMBER to replace the BCD sentence numbers with the binary equivalent.

NICETYPE - This subroutine can be used to print the phrase dictionary.

MERGER - This subroutine compares two strings to determine which of the two occurs first alphabetically.

SORT1 - This subroutine is a specialized variable record length alphabetic sort.

MERGE2 - This subroutine is a specialized alphabetic merge routine which is used by SORT1 to alphabetically sort an input file.

COMPARE - This subroutine compares two strings character for character and returns with the number of characters which match.

MESSAGES TO OPERATOR SUBROUTINES

CHOICE - This subroutine which is called when an end-of-file is read asks the operator if there is more input or if output is to be terminated. If there is more input, the operator is instructed to mount a specified tape.

GIVEMES - This subroutine informs the operator on which units input and output tapes are to be mounted.

MESG1 - This subroutine tells the operator to mount several tapes on the tape units in preparation for the beginning of Phrase X, where X = 1,2,3,4.

MESG2 - This subroutine requests operator to remove and mount specified tapes on the tape units at the end of Phrase X, where X = 1,2,3,4 in preparation for a CO-OP sort or at the end of Phrase 4, for a specialized sort.

SPECMESI - This subroutine requests operator to mount a tape on a specified tape unit.

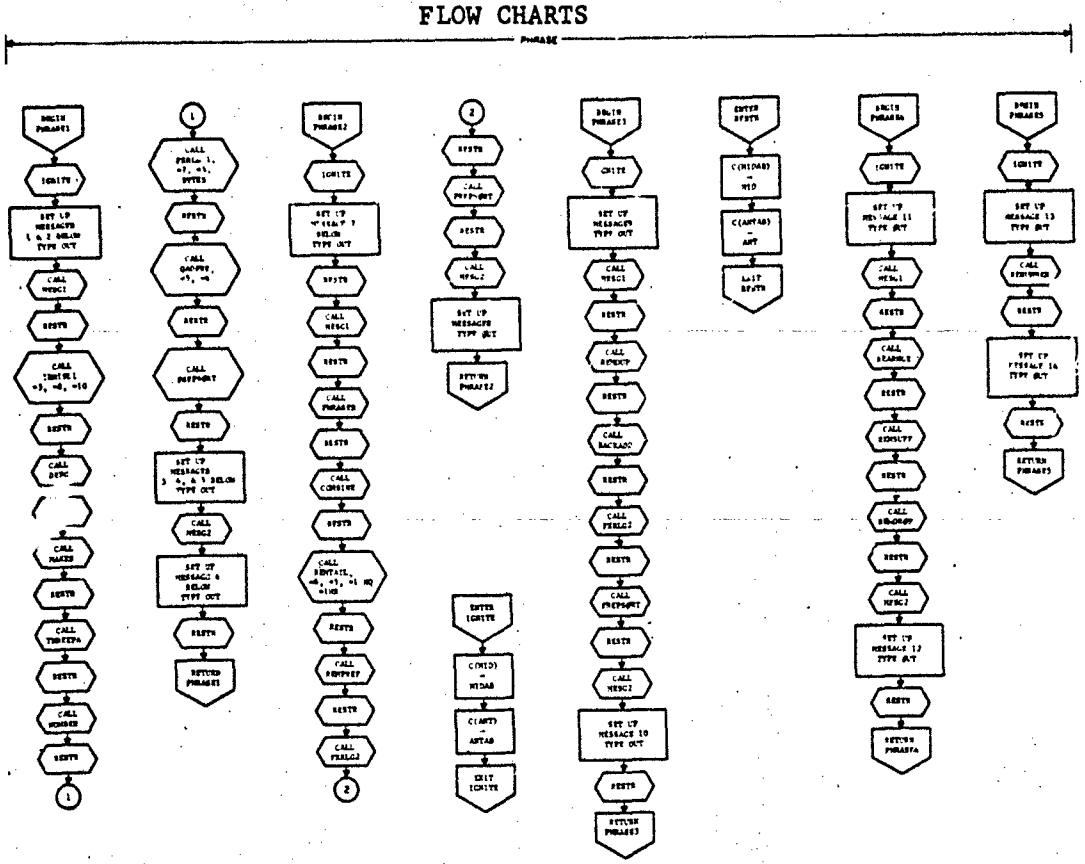
SPECMESO - This subroutine requests operator to remove and label a tape from a specified tape unit.

SPECWRIT - This subroutine calls the ISL write routine to write the specified record. On an end-of-tape condition, the operator is requested to remove and label the full tape and mount a new tape.

TONE1 - This subroutine generates an audible tone indicating man-machine interaction is required.

TYPEMES - This subroutine will type a string message to the operator on the console typewriter.

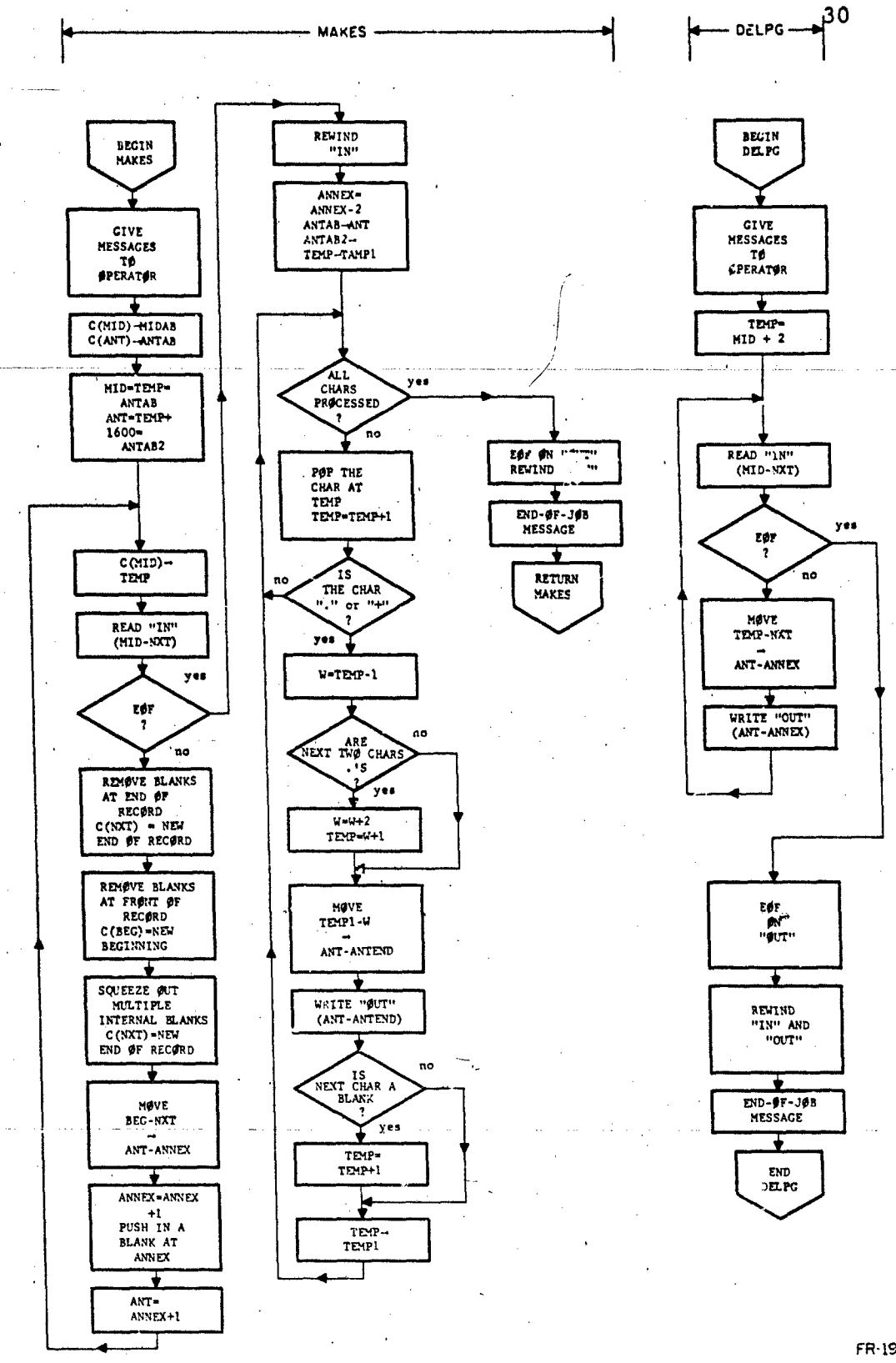
FLOW CHARTS



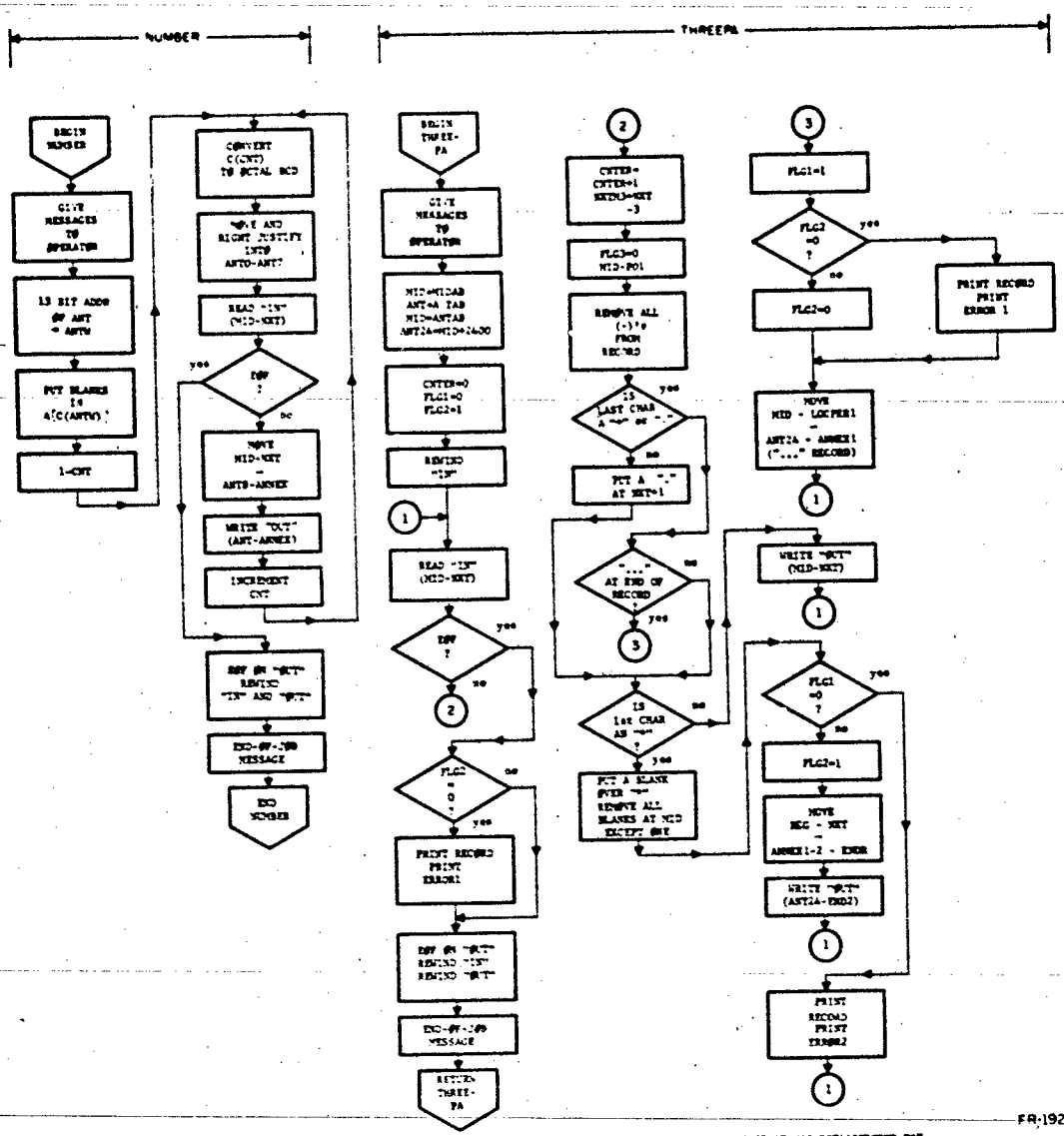
FR-1903

MESSAGES:

1 BEGIN PHRASE1
2 MOUNT FULL TAPE ON L.U. 8
3 REMOVE TAPE REEL FROM L.U. 7
4 LABEL TAPE REEL FROM L.U. 8 WITH L1
5 REMOVE TAPE REEL FROM L.U. 5 AND PUT AWAY
6 END PHRASE1
7 BEGIN PHRASE2
8 END PHRASE2
9 BEGIN PHRASE3
10 END PHRASE3
11 BEGIN PHRASE4
12 END PHRASE4
13 BEGIN PHRASE5
14 END PHRASE5


MAKES
DELPG
BEGIN MAKES
GIVE MESSAGES TØ ØPERATØR
C(MID)→MIDAB
C(ANT)→ANTAB
MID=TEMP=ANTAB
ANT=TEMP+1600=ANTAB2
C(MID)→TEMP
READ "IN" (MID-NXT)
EØF ?
yes
no
REMØVE BLANKS AT END ØF RECØRD
C(NXT) = NEW END ØF RECØRD
REMØVE BLANKS AT FRØNT ØF RECØRD
C(BEG)=NEW BEGINNING
SQUEEZE ØUT MULTIPLE INTERNAL BLANKS
C(NXT)=NEW END ØF RECØRD
MØVE BEG-NXT → ANT-ANNEX
ANNEX=ANNEX+1
PUSH IN A BLANK AT ANNEX
ANT=ANNEX+1
REWIND "IN"
ANNEX=ANNEX-2
ANTAB→ANT
ANTAB2→TEMP→TAMP1
ALL CHARS PRØCESSED ?
yes
no
PØP THE CHAR AT TEMP
TEMP=TEMP+1
IS THE CHAR "." or "+" ?
no
yes
W=TEMP-1
ARE NEXT TWØ CHARS .'S ?
no
yes
W=W+2
TEMP=W+1
MØVE TEMP1-W → ANT-ANTEND
WRITE "ØUT" (ANT-ANTEND)
IS NEXT CHAR A BLANK ?
no
yes
TEMP=TEMP+1
TEMP→TEMP1
EØF ØN "IN" REWIND "IN"
END-ØF-JØB MESSAGE
RETURN MAKES
BEGIN DELPG
GIVE MESSAGES TØ ØPERATØR
TEMP=MID + 2
READ "IN" (MID-NXT)
EØF ?
yes
no
MØVE TEMP-NXT → ANT-ANNEX
WRITE "OUT" (ANT-ANNEX)
EØF ØN "ØUT"
REWIND "IN" AND "OUT"
END-ØF-JØB MESSAGE
END DELPG

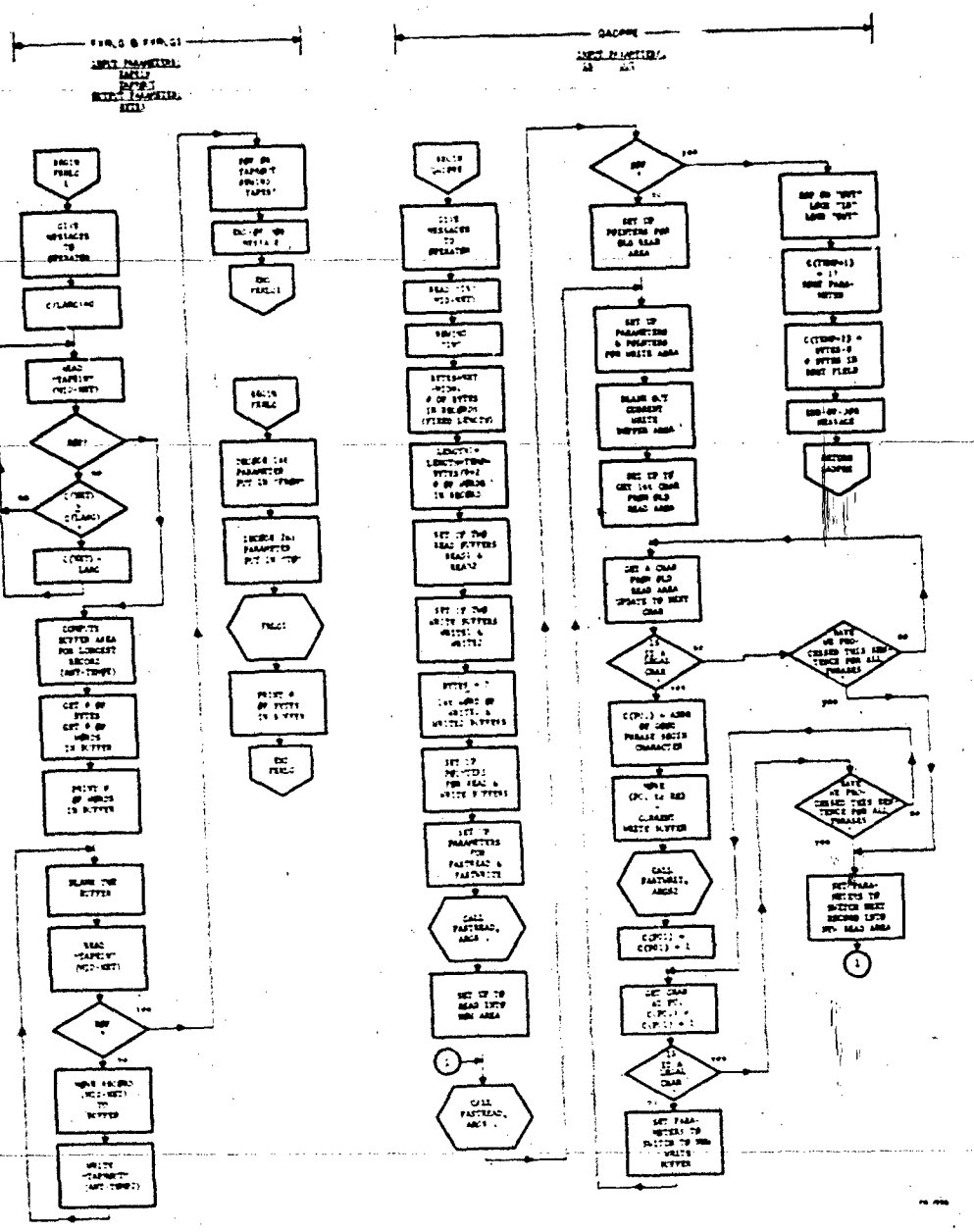
FR-1904

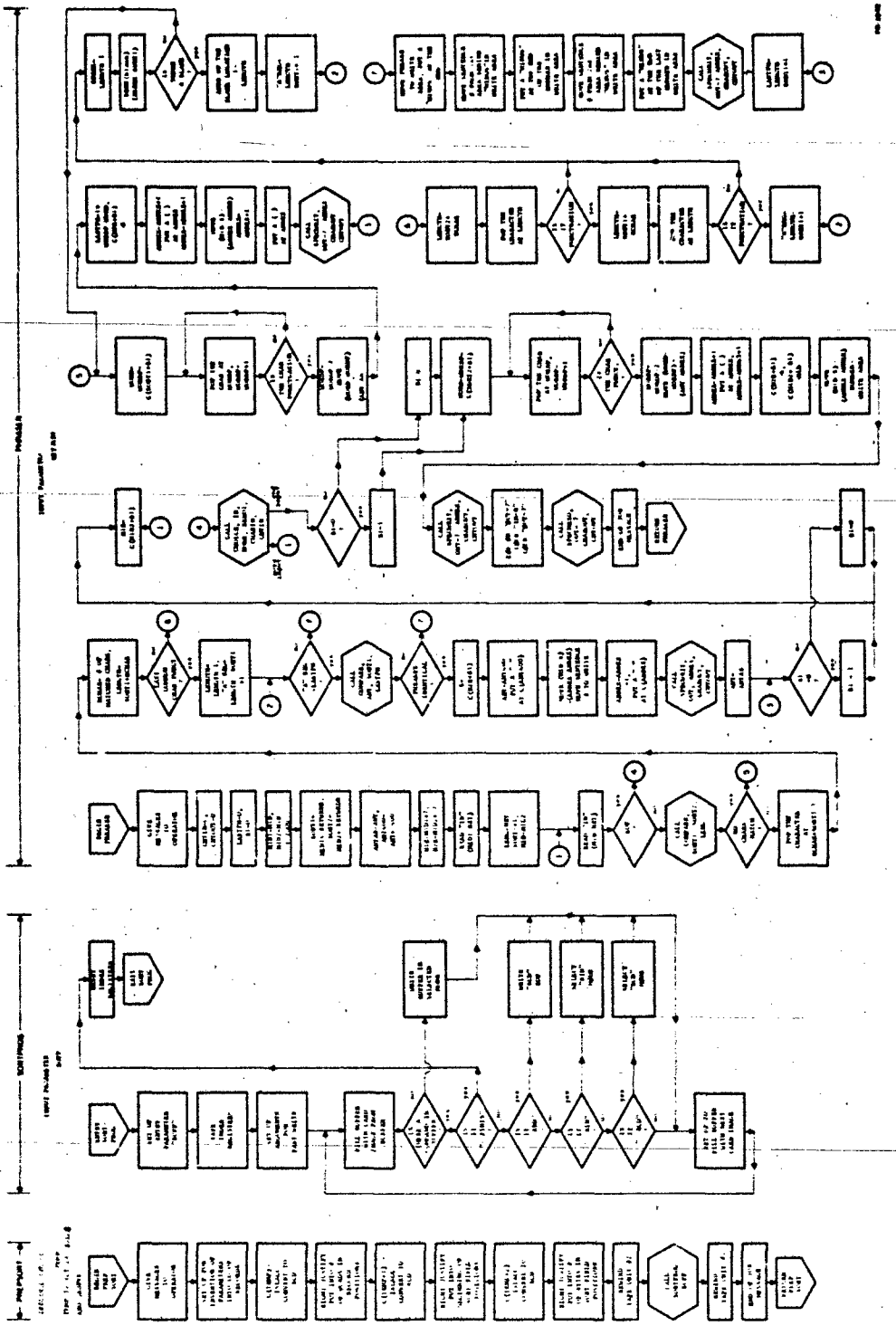


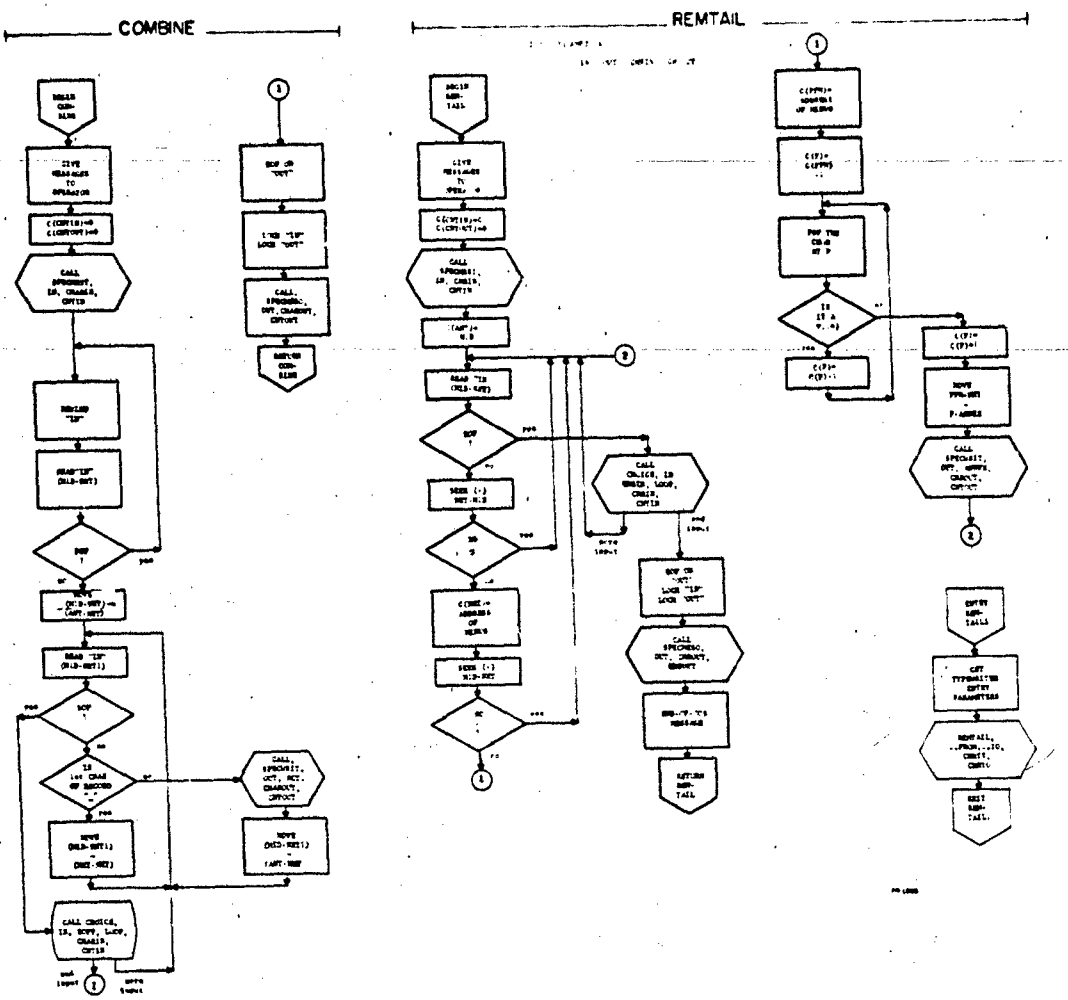
FA-1926

ERROR1 "THE FOLLOWING SENTENCE WAS NOT APPENDED TO"

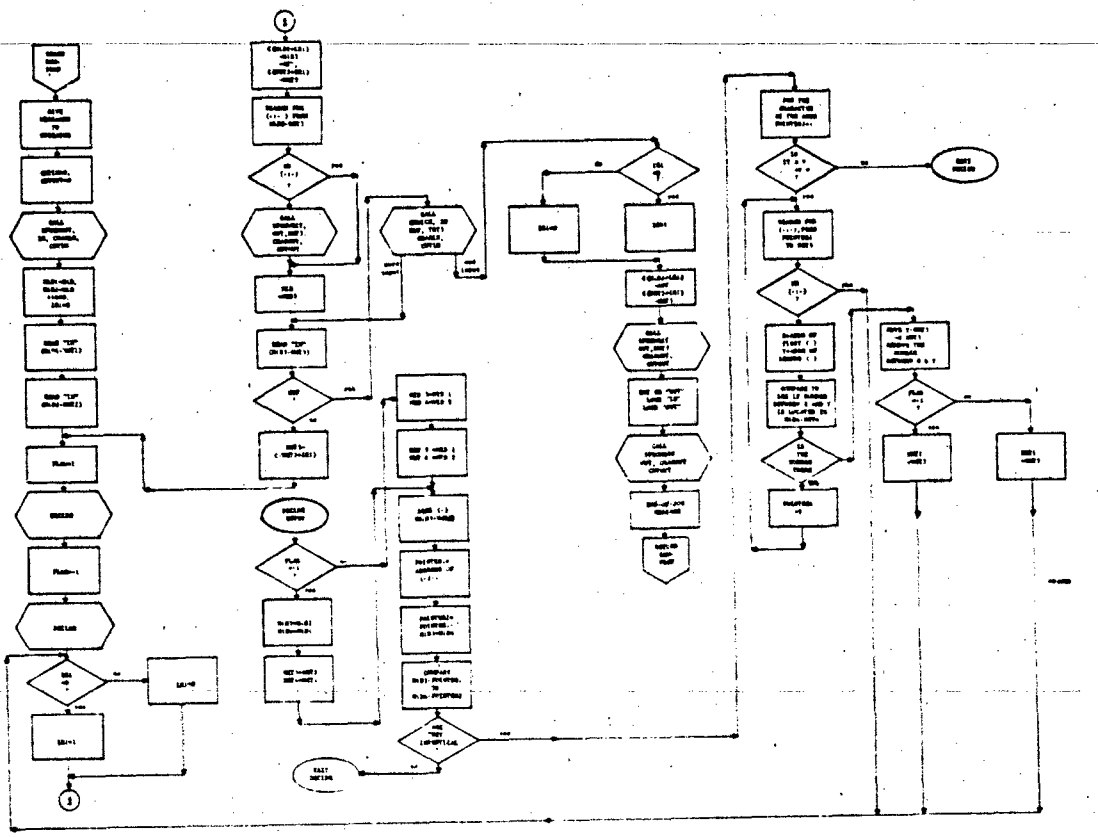
ERROR2 "THERE WAS NO SENTENCE WITH ... TO APPEND TO"



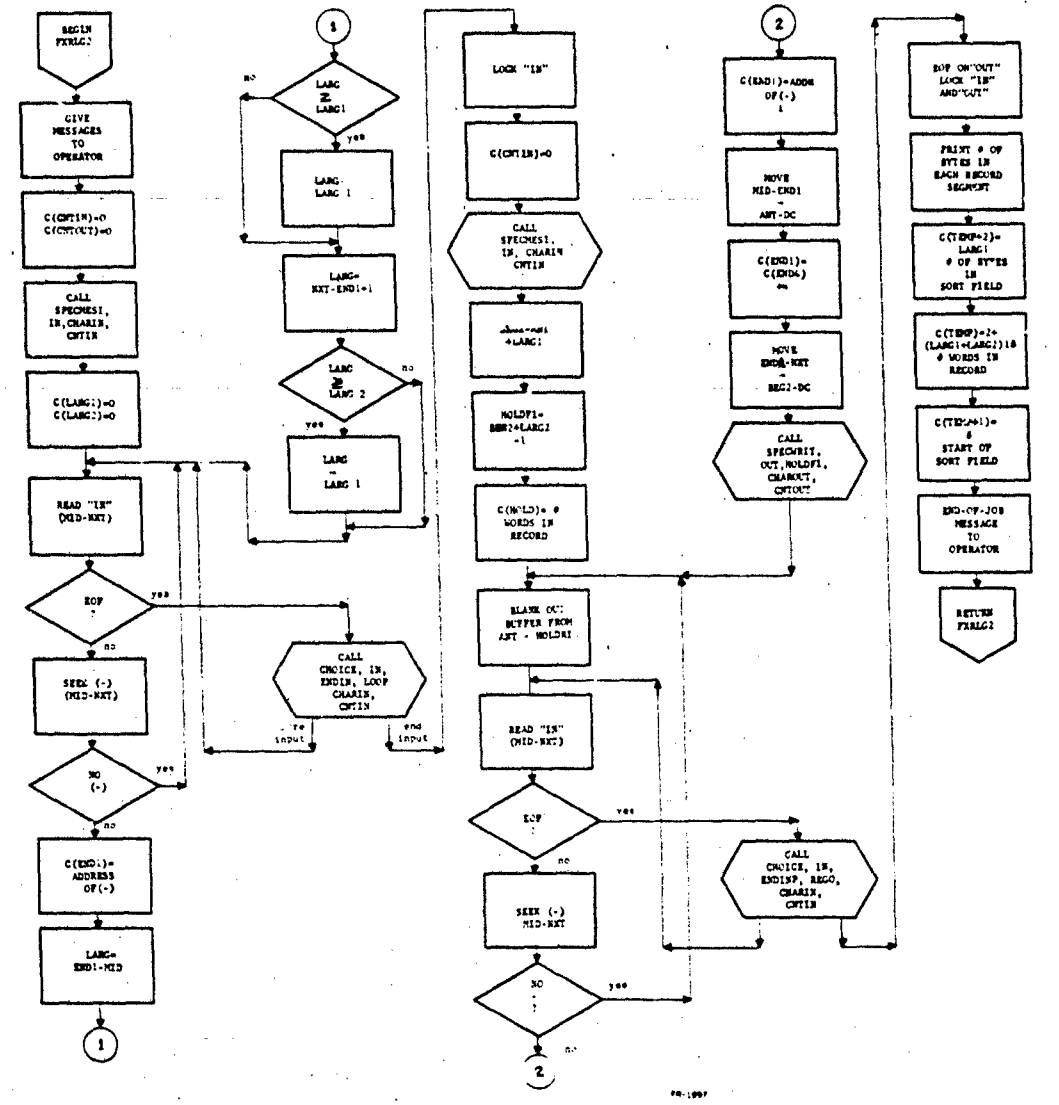



COMBINE
REMTAIL

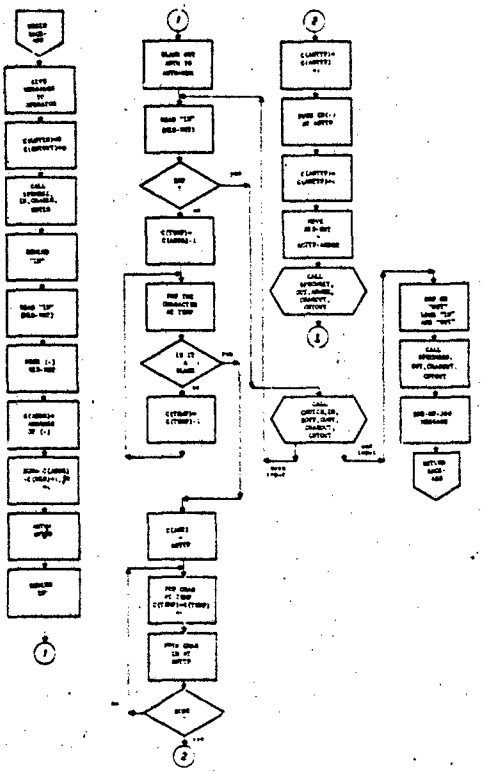
REMPREF



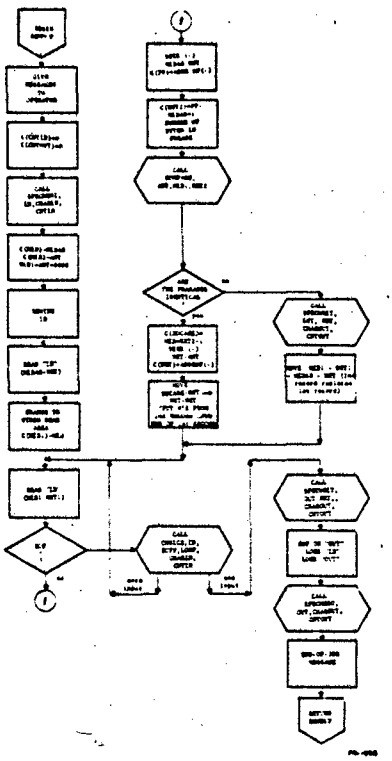
FXRLG2

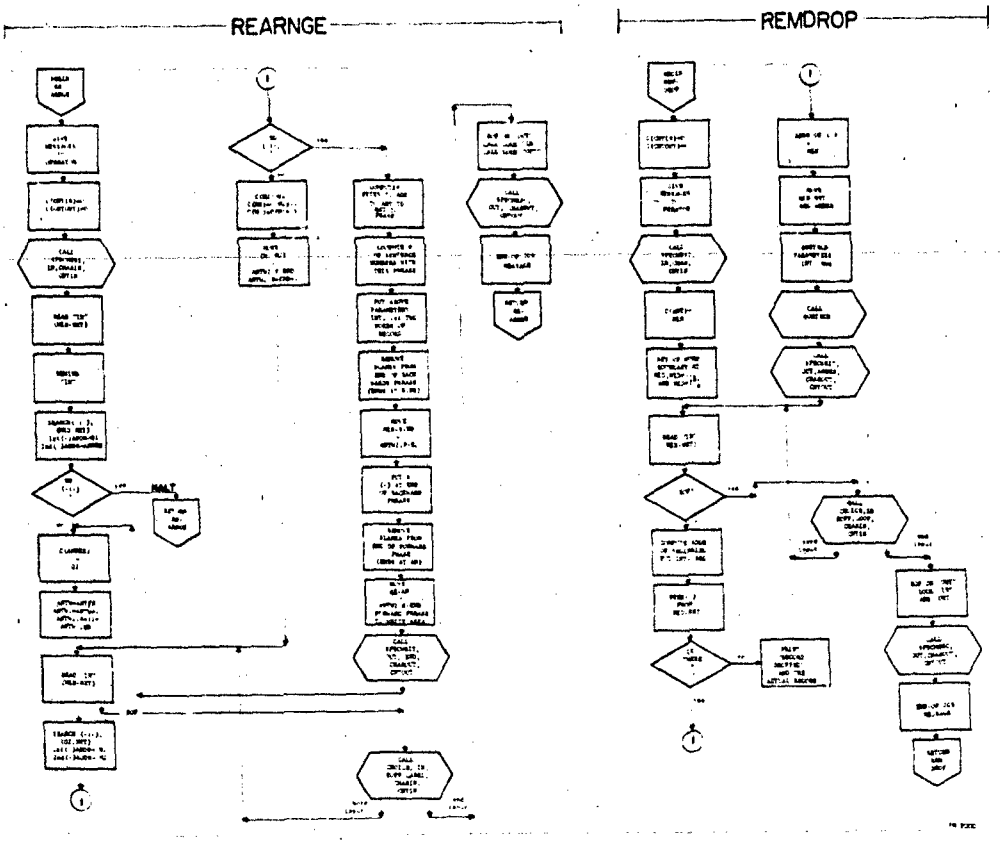

BEGIN FXRLG2
GIVE MESSAGES TO OPERATOR
C(CNTIN)=0 C(CNTOUT)=0
CALL SPECMESI, IN,CHAIN, CNTIN
C(LARG1)=0 C(LARG2)=0
READ "IN" (MID-NXT)
EOF
yes
no
SEEK (-) (MID-NXT)
NO (-)
yes
no
C(END1)= ADDRESS OF(-)
LARG= END1-MID
1
1
LARG ≥ LARG1
no
yes
LARG= LARG 1
LARG= NXT-END1+1
LARG ≥ LARG 2
no
yes
LARG = LARG 1
CALL CHOICE, IN, ENDIN, LOOP CHAIN, CNTIN
no input
end input
LOCK "IN"
C(CNTIN)=0
CALL SPECMESI, IN, CHAIN CNTIN
+LARG1
HOLDFI= BBR2+LARG1 -1
C(WORLD)= # WORDS IN RECORD
BLANK OUT BUFFER FROM ANT - HOLDRI
READ "IN" (MID-NXT)
EOF
yes
no
SEEK (-) MID-NXT
NO
yes
no
2
2
C(END1)=ADDR OF(-) 1
MOVE MID-END1 = ANT-DC
C(END1)= C(END4)
MOVE END4-NXT = BEG2-DC
CALL SPECWRIT, OUT,HOLDFI, CHAROUT, CNTOUT
CALL CHOICE, IN, ENDISP, REGO, CHAIN, CNTIN
EOF ON "OUT" LOCK "IN" AND "OUT"
PRINT # OF BYTES IN EACH RECORD SEGMENT
C(TEMP+2)= LARG1 # OF BYTES IN SORT FIELD
C(TEMP)=2+ (LARG1+LARG2)/8 # WORDS IN RECORD
C(TEMP+1)= 4 START OF SORT FIELD
END-OF-JOB MESSAGE TO OPERATOR
RETURN FXRLG2

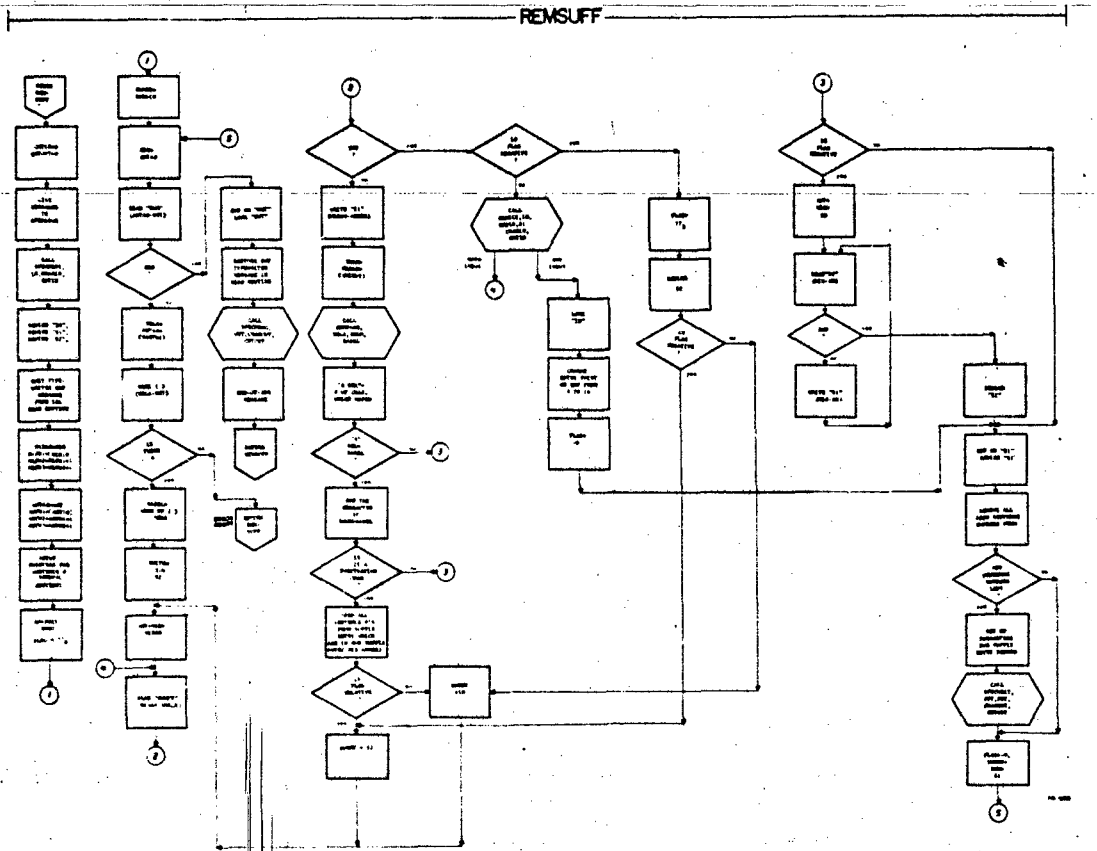
BACKADD



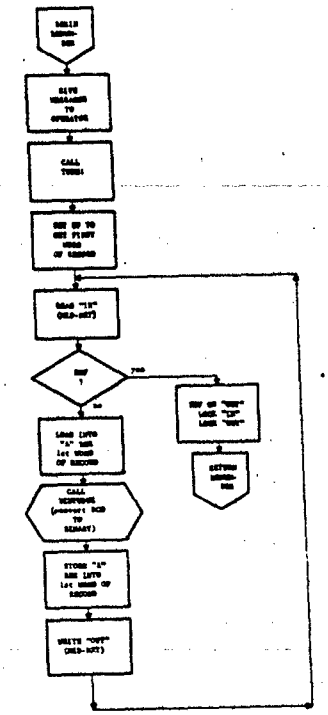
REMDUP



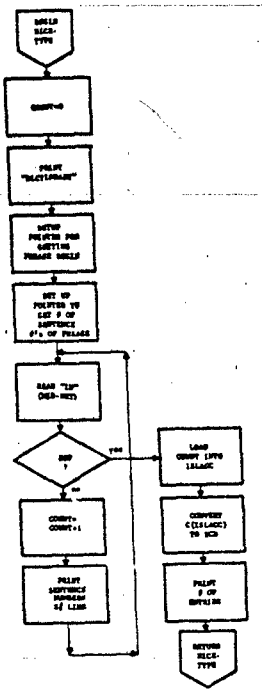

REARNGE
REMDROP
HALT


REMSUFF

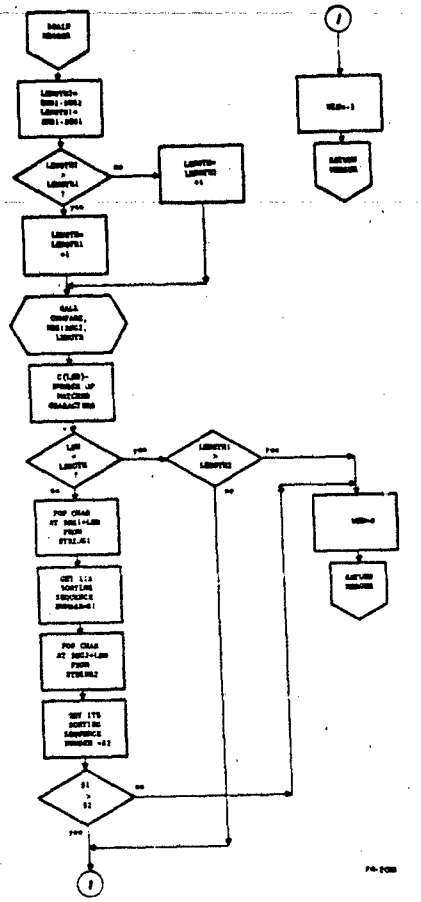
RENUMBER

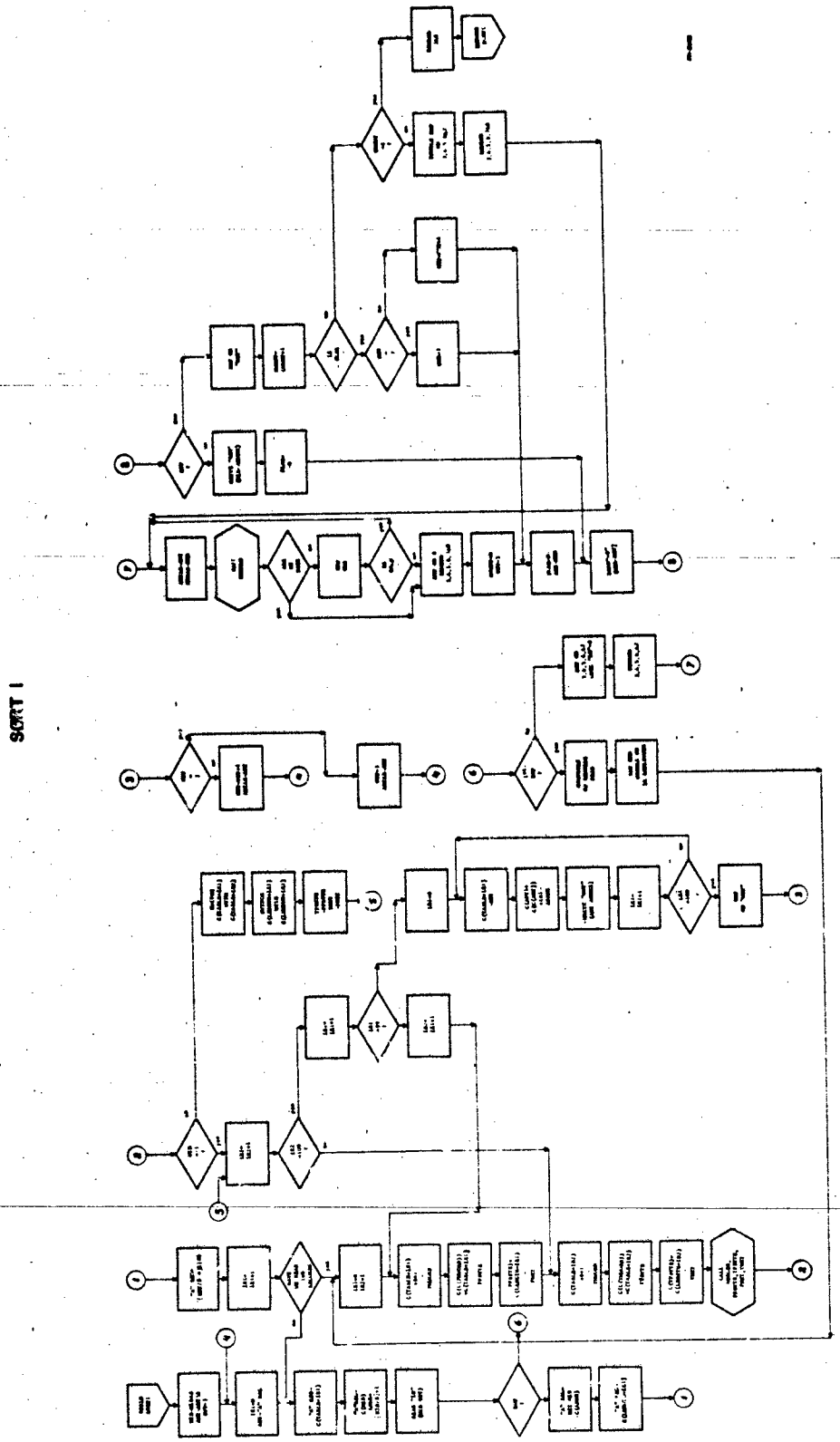


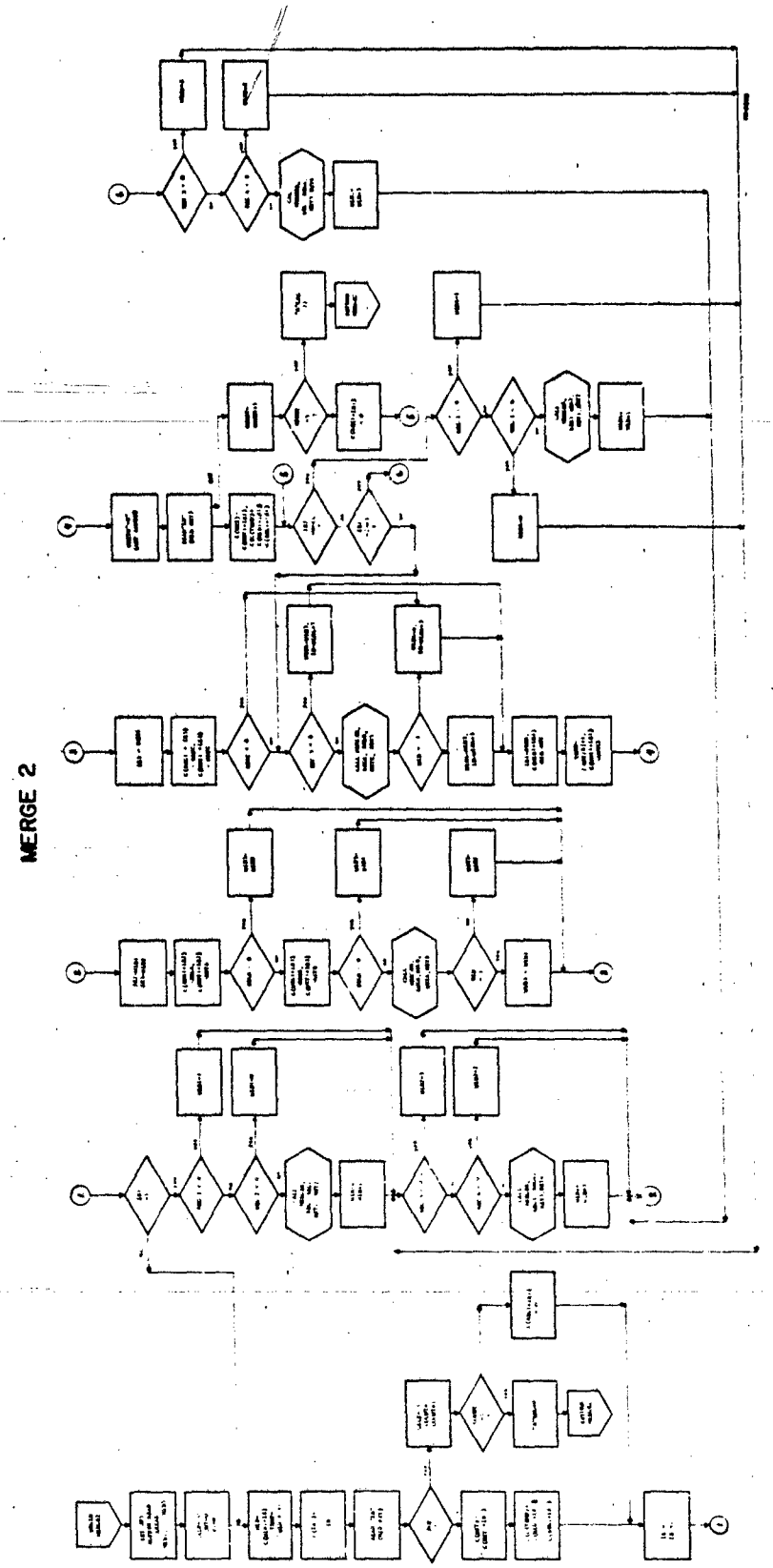
NICETYPE



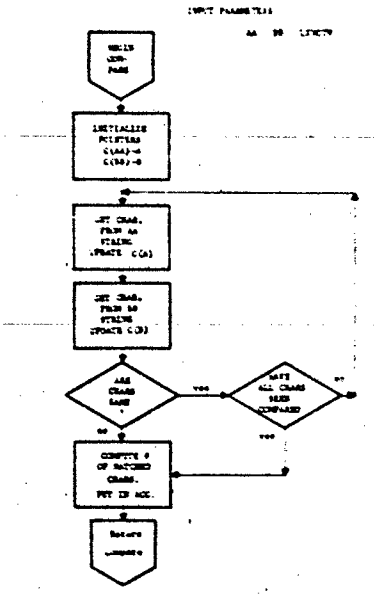
MERGER



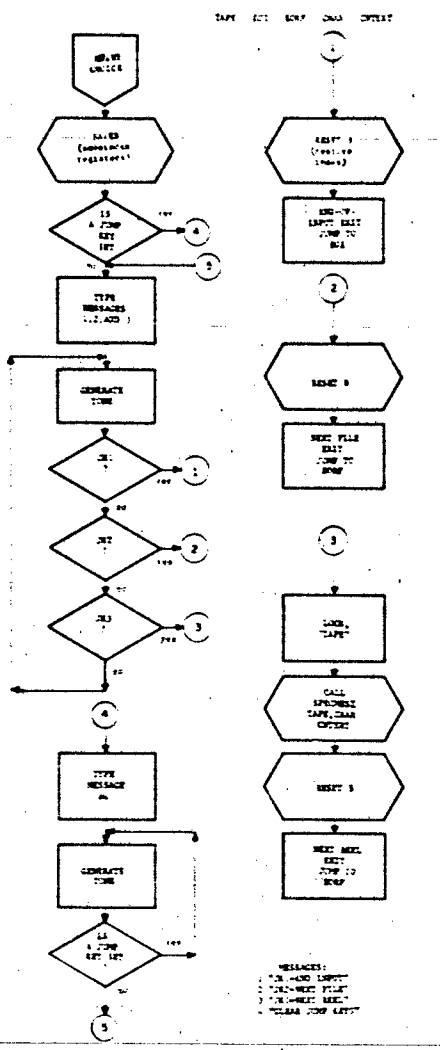

SORT I


MERGE 2

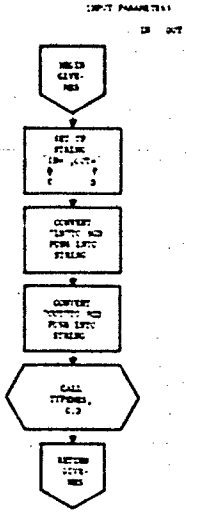
COMPARE



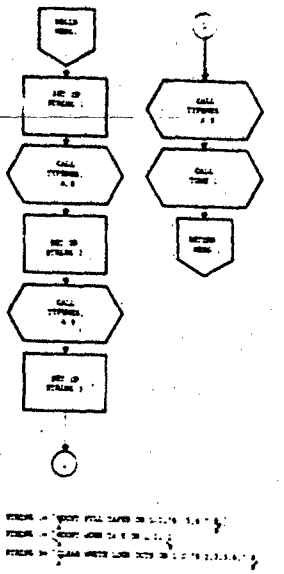
CHOICE



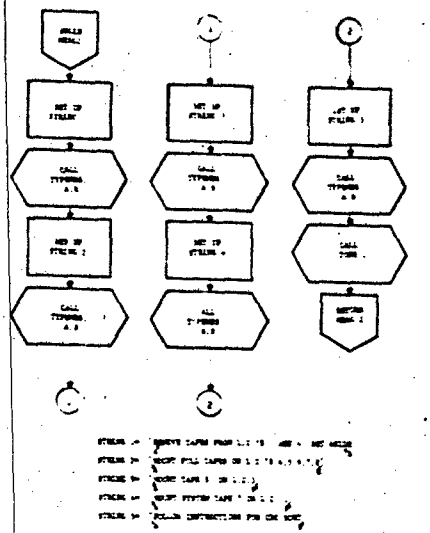
GIVEMES



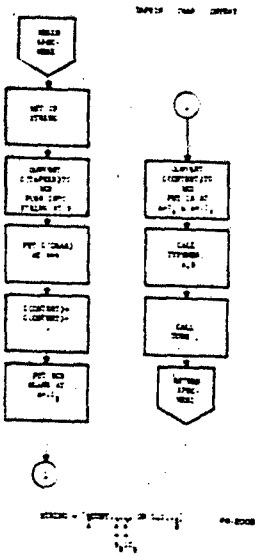
MESG 1

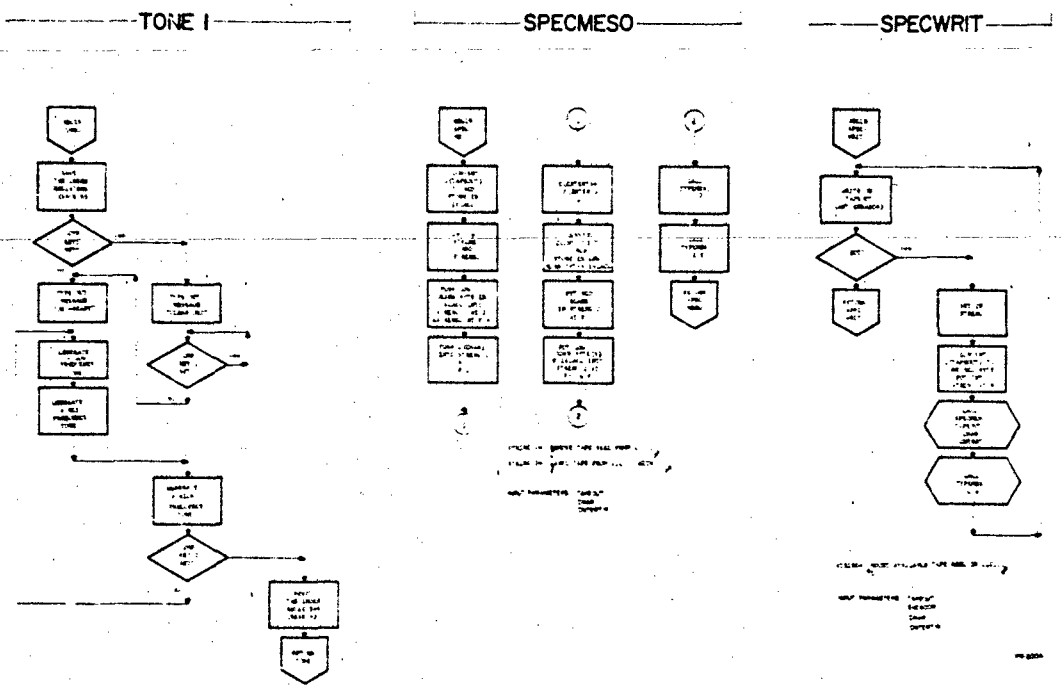


MESG 2

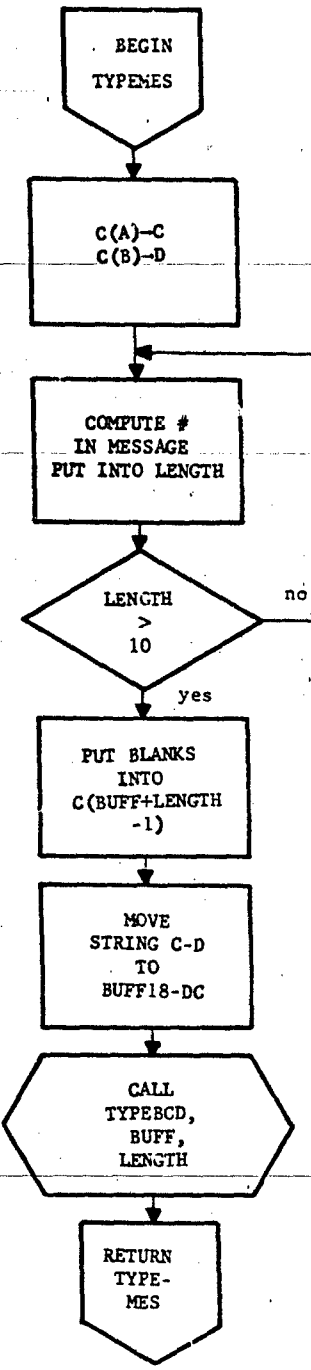


SPECMESI




TONE I
SPECMESO
SPECWRIT

TYPEMES



INPUT PARAMETERS: A

B

MESSAGE: (error type)

STRING=
"TYPE MESSAGE TOO LONG"
C D

FR-2007

INSTRUCTIONS FOR THE CONSTRUCTION OF THE PHRASE DICTIONARY

1. Mount tape C-140 (ISL Master) on logical unit 1.
2. Mount tape C-99 (Dictionary Construction Programs) on logical unit 4.
3. Mount the edited source input tape on logical unit 3.
4. Mount a scratch tape on logical unit 2.
5. Press "Autoload."
6. When the typewriter returns with the sign"#", type: "load,4 cr" where "cr" indicates a carriage return.
7. When the typewriter again returns with "#", type: "go, phrase1 cr".
8. Follow the instructions given on the console typewriter.
9. When "end phrase1" appears, mount the tape labelled S1 on logical unit 3, remove all other tapes, except the scratch tape on logical unit 2.
10. Mount tape 7 on logical unit 1 and press "Autoload."
11. When the "*" returns on the typewriter, type "cr". When the "." returns, type "m.cr". Follow the instructions given on the typewriter. The SORT/REFERENCE manual can be used for the methods of treating errors in this step.
12. When the job is complete, mount the output tapes (label as T1,T2,...) on logical unit 8.
13. Mount C-140 and C-99 as in steps 1 and 2 above.
14. Follow the instructions given in steps 5,6,7,8,9,10,11,12, changing step 7 each time to "go, phrase2", "go,phrase3", and finally "go, phrase4"

15. When the typewriter returns with "#" after the statement "end phrase4" type: "call,4,sort1 cr".
16. Mount the tape labelled S1 on logical unit 8.
17. Type: "go,sort1 cr".
18. Remove the output tape from logical unit 3 -- this is the phrase dictionary.
19. Type: "go,phrase5 cr" after the "#" appears.
20. The edited data base is on logical unit 7 when the "#" reappears.
21. Label the tapes from steps 18 and 20 with appropriate labels and remove write rings.

APPENDIX B

INSTRUCTIONS FOR THE INITIAL STRATEGIES

Instructions for the operation of the strategies are given below.

1. Mount tape C-140 on logical unit 1.
2. Press "Autoload".
3. Mount D-99 on logical unit 2 and mount D-98 on logical unit 3.
4. Type: "load,2,3(cr)".
5. After the "#" sign, type: go,runallst(cr)".
6. State a question followed by a "%" sign.
7. Give a strategy using the following code

$$A_1B_1C_1D_1A_2B_2C_2D_2\ldots A_6B_6C_6D_6\%$$

where

A_i = 1,2,3 for strategy 1, strategy 2, or strategy 3, respectively.

B_i = N or Z, N representing that the second strategy option "ENS" will be exercides, or Z, it will not be exercised.

C_i = 1 or 2, 2 indicates that the threshold of 2 option will be exercised.

D_i = space

The % sign must follow the last $A_iB_iC_iD_i$, sequence which can be any length, i.e., i = 1,2,...,6.

8. By typing "SAME" when the word "Question" is typed, the previous question which was entered will be used.
9. To terminate use of the R2 system, type "ABORT" when the word "QUESTION" is typed.

APPENDIX C

A PORTION OF THE PHRASE DICTIONARY

A sample of the "phrase dictionary" for the R2 system with "Rules of the Road" as the data base is shown below.

DRIVERS ARE PROHIBITED FROM - 646,686,739

DRIVERS ARE REQUIRES - 378,1019

DRIVERS LICENSE - 33

DRIVERS MAY - 532,777

DRIVERS MUST - 577,668,689,1011

DRIVERS MUST ALWAYS - 569,579

DRIVERS MUST FIRST - 555,573

DRIVERS MUST YIELD THE RIGHT OF WAY TO - 571,609

DRIVERS MUST YIELD TO - 607,581

DRIVERS MUST YIELD TO PEDESTRIANS - 567,575

DRIVERS OF - 450,557,560,1073,1299

DRIVERS TO - 404,413,634,687,1023

DRIVERS WHO - 168,1164

DRIVERS WHO ARE - 149,605

DRIVERS WILL BE - 34,35

DRIVES - 1189,1211,1284

DRIVEWAY - 837,837

DRIVEWAYS - 808

DRIVING - 61,61,77,152,203,208,209,283,453,453,594,645,949,949,1212,1212

DRIVING A - 101,106

DRIVING ABILITY - 1172,1172,1173,1195,1195,1200,1208

APPENDIX D

EXAMPLES OF THE STRATEGIES

Examples of the strategies are shown below.

Examples 1, 2, and 3 show three different strategies applied to the question

"An octagon sign always means to stop?"

The abbreviations shown following the questions indicate the following:

ST1 = strategy 1

ST2 = strategy 2

ST3 = strategy 3

TH1 = threshold of two option not exercised

TH2 = threshold of two option exercised

N-1, N, N+1 = ENS option exercised

The asterisks to the left of a sentence indicate those retrieved statements which are relevant.

Examples 4 and 5 serve to indicate the need for the second relevancy rate, R_2. In example 5, there are two relevant answers retrieved. These, in fact, are the only two relevant answers contained in the data base. With R_1, the strategy used in Example 4 is not penalized for having retrieved only one of the two relevant answers. With R_2, the stragety is appropriately penalized.

AN OCTAGON SIGN ALWAYS MEANS TO STOP

ST1, TH2

*THE EIGHT SIDED (OCTAGON) SHAPE ALWAYS MEANS STOP.

A ROUND SIGN ALWAYS MEANS RAILROAD CROSSING.

THIS SIGN TELLS YOU TO LOOK, LISTEN AND SLOW DOWN BECAUSE YOU MAY HAVE TO STOP.

IF A TRAIN IS APPROACHING STOP,

STOP,

PASSING IS PROHIBITED ON A TWO LANE HIGHWAY IN A NO PASSING AREA MARKED BY A YELLOW STRIPE OR A DO NOT PASS SIGN.

RAILROAD CROSSINGS ARE ILLEGAL WHEN A FLAGMAN IS GIVING A SIGNAL TO STOP.

EXAMPLE 1

AN OCTAGON SIGN ALWAYS MEANS TO STOP

ST2, TH2

*THE EIGHT SIDED (OCTAGON) SHAPE ALWAYS MEANS STOP.

A ROUND SIGN ALWAYS MEANS RAILROAD CROSSING.

THIS SIGN TELLS YOU TO LOOK, LISTEN AND SLOW DOWN BECAUSE YOU MAY HAVE TO STOP.

PASSING IS PROHIBITED ON A TWO LANE HIGHWAY IN A NO PASSING AREA MARKED BY A YELLOW STRIPE OR A DO NOT PASS SIGN.

RAILROAD CROSSINGS ARE ILLEGAL WHEN A FLAGMAN IS GIVING A SIGNAL TO STOP.

$$R_1 = R_2 = \frac{1}{5}$$

EXAMPLE 2

AN OCTAGON SIGN ALWAYS MEANS TO STOP

ST3, TH1

*THE EIGHT SIDED (OCTAGON)SHAPE ALWAYS MEANS STOP.

A ROUND SIGN ALWAYS MEANS RAILROAD CROSSING.

$R_1 = R_2 = \frac{1}{2}$

EXAMPLE 3

IS IT LEGAL TO CROSS A DOUBLE YELLOW LINE TO COMPLETE A PASS+

ST1, TH2

*CROSSING A DOUBLE YELLOW IS PROHIBITED EXCEPT WHEN TURNING INTO A DRIVEWAY.

$R_1 = 1\ R_2 = \frac{1}{2}$

EXAMPLE 4

IS IT LEGAL TO CROSS A DOUBLE YELLOW LINE TO COMPLETE A PASS+

ST2, TH1

SERVICE PERSONNEL OF THE ARMED FORCES WHO ARE LEGAL RESIDENTS OF ILLINOIS ARE GIVEN A 30 DAY GRACE PERIOD FOLLOWING THEIR RETURN TO THE CONTINENTAL LIMITS OF THE UNITED STATES.

TWO REQUIRED (ONE ON MOTORCYCLES), WITH WHITE OR YELLOW TINTED LIGHTS VISIBLE FOR AT LEAST 500 FEET.

TWO YELLOW LIGHTS, ONE ON EACH UPPER FRONT CORNER, VISIBLE FOR 500 FEET.

ONE YELLOW REFLECTOR ON THE LOWER LEFT HAND AND RIGHT HAND FRONT CORNERS.

THIS ONE IS USED ON TWO LANE ROADS AND IT WARNS DRIVERS THAT THEY ARE APPROACHING AN AREA WHERE TWO CARS CANNOT SAFELY PASS EACH OTHER WITHOUT SLOWING DOWN.

THERE ARE THREE TYPES FOUR OR MORE LANES, DOUBLE SOLID YELLOW STRIPES.

*CROSSING A DOUBLE YELLOW IS PROHIBITED EXCEPT WHEN TURNING INTO A DRIVEWAY.

*DRIVERS MAY CROSS THE YELLOW STRIPE TO COMPLETE A PASSING MANEUVER WHICH WAS STARTED BEFORE THE BEGINNING OF A NO PASSING ZONE, AND THE NO PASSING STRIPE MAY ALSO BE CROSSED WHEN MAKING A LEFT TURN INTO A DRIVEWAY.

A YELLOW LINE EXTENDS THE ENTIRE DISTANCE TO PREVENT CROSSING OVER, AND TWO WHITE STRIPES ARE PAINTED ON EACH SIDE OF THE TRACK.

IN URBAN AREAS THE LINE IS USUALLY PLACED FROM THREE TO FIVE FEET IN ADVANCE OF THE CROSSWALK.

YELLOW LINES AND WHITE DIAGONAL STRIPES ARE USED TO MARK FIXED OBSTRUCTIONS OR TO CHANNEL TRAFFIC INTO DIFFERENT LANES.

THEY ARE DEVELOPED THROUGH CAREFUL STUDY BY TRAFFIC ENGINEERS, SAFETY EXPERTS AND LEGAL AUTHORITIES, AND ONLY BY STRICT OBSERVANCE CAN WE EXPECT TO TRAVEL WITH A REASONABLE DEGREE OF SAFETY AND SPEED ON OUR CROWDED, COMPLEX STREETS AND HIGHWAYS.

IF HE IS WITHIN AN INTERSECTION CONTROLLED BY A SIGNAL LIGHT HE MAY COMPLETE HIS TURN WHEN THE LIGHT TURNS RED, BUT IF HE HAS NOT YET ENTERED THE INTERSECTION HE MUST WAIT FOR THE NEXT GREEN LIGHT.

PASSING IS PROHIBITED ON A TWO LANE HIGHWAY IN A NO PASSING AREA MARKED BY A YELLOW STRIPE OR A DO NOT PASS SIGN.

IF YOU PASS THE SAME SPOT BEFORE YOU ARE THROUGH COUNTING YOU ARE FOLLOWING TOO CLOSELY.

ELECTRICAL TURN SIGNALS AND THE FLASHING OF ELECTRIC TURN SIGNALS AS A COURTESY OR DO PASS INDICATION TO OTHER DRIVERS IN THE REAR IS PROHIBITED BY LAW.

IT IS NOT NECESSARY TO STOP FOR A SCHOOL BUS ON A CONTROLLED ACCESS HIGHWAY IF THE BUS IS IN A LOADING ZONE AT A PLACE WHERE PEDESTRIANS ARE NOT PERMITTED TO CROSS.

USE THE WHITE STRIPE AT THE EDGE OF THE PAVEMENT AS A GUIDE LINE.

THIS DESIGN FEATURE ENABLES VEHICLES TO CROSS, ENTER OR LEAVE EITHER HIGHWAY WITHOUT INTERFERING WITH OTHER VEHICLES.

WHEN THE WEATHER IS BAD OR THE PAVEMENT IS SLICK, DOUBLE OR TRIPLE THE INTERVAL.

A NEW INEXPERIENCED PASSENGER SHOULD BE INSTRUCTED AS FOLLOWS THAT RAILROAD TRACKS OR STEEL BRIDGE EXPANSION JOINTS PRESENT A SPECIAL PROBLEM, TRY TO CROSS THEM AT AN ANGLE, IF POSSIBLE, TO AVOID SKIDDING OR CATCHING THE WHEELS.

THESE ARE AREAS FOR COMPLETE HONESTY BETWEEN THE PHYSICIAN AND PATIENT.

$$R_1 = \frac{1}{11} \qquad R_2 = \frac{1}{11}$$

EXAMPLE 5

Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author)
University of Illinois
Coordinated Science Laboratory
Urbana, Illinois 61801

2a. REPORT SECURITY CLASSIFICATION

2b. GROUP

3. REPORT TITLE

PHRASE DICTIONARY CONSTRUCTION METHODS FOR THE R2 INFORMATION RETRIEVAL SYSTEM

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)

5. AUTHOR(S) (First name, middle initial, last name)

JANSEN, James Merritt, Jr.

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
December, 1969	58	4

8a. CONTRACT OR GRANT NO.
DAAB 07-67-C-0199; also in part
b. PROJECT NO. OE C-1-7-071213-4557.
c.
d.

9a. ORIGINATOR'S REPORT NUMBER(S)
R-447

9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)

10. DISTRIBUTION STATEMENT

This document has been approved for public release and sale; its distribution is unlimited.

11. SUPPLEMENTARY NOTES

12. SPONSORING MILITARY ACTIVITY
Joint Services Electronics Program
thru U. S. Army Electronics Command
Fort Monmouth, New Jersey 07703

13. ABSTRACT

A set of programs that receives questions and a data base in natural language attempts to recall those statements in the data base that could best be used in answering the question is described. A variety of strategy techniques are allowed.

DD FORM 1 NOV 65 1473

Security Classification

Security Classification

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Information Retrieval						
Question Answering						
Phrase Dictionary						
Maximal Phrase						

Security Classification