V. THE DICTIONARY LOOK-UP SYSTEM

M. Cane

1. Introduction

The first link of the SMART system converts unprocessed English text into vectors representing their documents. The vectors consist of concept numbers included in the thesaurus, and of concept numbers from the cluster dictionary. The source of each concept number in the vector is noted in the tag field of the corresponding vector entry. The tag field identifies concepts as originating in the normal synonym dictionary (thesaurus), or in the statistical phrase dictionary; concepts are also identified as coming from the body of the text or from the title of a document (that is, the first sentence of the document). The completed document vectors are written onto a tape (A5) in binary form. Each document will consist of at least two tape records. The first record is twelve words long, and contains the image of the title card for the document; following this, the vector entries are written onto the tape in 260 word records. The vector is terminated with a word of zeros.

In addition to this vector tape the first link may produce a number of other tapes. If the option ENGTXT is operative, the English text of the document is written on the system output tape (A3). If NOTFND is used, information concerning the words not found in the dictionary is written on tapes B1 and B2, so that it may be processed by the second link. If the SYNTAX option is operative, sentences which were found to contain statistical phrases are written on tape A7 for subsequent syntactic processing.

If PUNCH is on, or ANSWER is greater than one, the title card (the "*" card) and any citations are written onto tape A4. If the document treated is identified by *FIND (the document is a search request) then the whole text is written onto tape (A4).

2. Program Description

The operations of the supervisory program for the first link are described in Flowchart I. First CAISET is called to set up the specifications. The operations of the CAISET program are detailed in Sec. IV of this report. The supervisor then rewinds the appropriate tapes and initializes the flags that are internal to the link. Processing may now begin. The supervisor reads a card from the system input tape (A2), and transfers according to the information appearing in columns 1-6 of this card. The second and third words (columns 7-18) of a *FILE card contain the name of the document collection to be processed. A *TIME card causes a time message to be written on tape A3. A *STOP card indicates that there are no further cards to be read from A2, and a *ONLY card indicates that no additional cards are to be read by the first link – all unprocessed texts must precede this card.

The major processing of the link is done when the "*" card indicates that a text follows. These cards must occur in the following order: first, all *LIST and *FIND documents which are assumed to be requests; a *FIND causes the whole text of the request to be written onto A4. Second, all *LIKE documents; these are treated both as requests, and as texts to be

considered a part of the document collection which is being searched. Last, all *TEXT documents; these are taken to be documents that are part of the collection to be searched. Documents which do not appear in the prescribed order are not processed.

The actual processing of each document begins at the point labelled "G" on the flowchart. The texts are processed one buffer load at a time, so that the first major task is to fill an array called IWDLST containing one thousand ten-word items, one item per word of input text. The first word of the item contains the sentence number and word-in-sentence number in the decrement and address, respectively. The next four words contain the BCD English word. These five items are put into IWDLST by subroutine SEGMNT. The last five words are entered into IWDLST by subroutine LOOK. The first three of these words contain up to six concept numbers packed into the decrement and address. If SYNTAX is off, the last two words are not filled. If it is on, these last two words will contain up to nine eight bit syntax codes, the last of which is a code from the suffix of the word (cf. Information Storage and Retrieval, Report No. ISR-7, Sec. IV).

SEGMNT returns control to the main program under two conditions. The first is that an entire document has been segmented into its component words, and the IWDLST buffer is not filled. In this case the supervisor transfers control to the point labelled "A" on the flowchart, thus initiating the processing of a new control card. The second condition is met when the buffer is filled. This may happen under the following circumstances:

1. Twenty complete documents are already included in this buffer.

2. After processing the last previous document, SEGMNT discover that there is room in IWDLST for only forty or fewer additional items. Since the next document may be expected to consist of more than 40 words, and since documents split between buffers are somewhat more expensive to process (see description of VECFM below, and of the statistical phrase routines (PHROCC) by G. Shapiro, elsewhere in this report) the buffer is considered filled at this point.

3. SEGMNT has used up the IWDLST buffer while attempting to process a document. In this case, a flag is set (named INCPLG) to tell other programs that the last document of this bufferload is incomplete. Normally SEGMNT returns the index in IWDLST of the last word of a document to a common location named NUMWRD. In the case where the document has not been completely processed, SEGMNT returns the index of the last word of the last complete sentence which was in part processed.

When any of these three conditions are met control passes to the point labelled "K" in the flowchart. Subroutines LOOK looks up each word included in the buffer load in a synonym dictionary (thesaurus), and returns concept numbers and syntactic codes to IWDLST. The algorithm used by LOOK is described in detail in Report No. ISR-7, Section IV. The present version of LOOK differs from the version described in Report No. ISR-7 in two significant ways:

1. In the earlier version it was required that the whole dictionary be located in core at the same time. The present version allows the dictionary tree to be broken down into pieces (see the description of the dictionary setup procedures, described elsewhere in this report). For each piece of the tree, the program scans through IWDLST, and searches for only those words which could be included in the part of the dictionary being treated at the moment. The read-in of the suffix and dictionary trees is handled by a set of subroutines called SREADL, TREAD1, and TREADL. This subroutine organization makes it possible to overlap the read of the first piece of the dictionary with the text processing performed by SEGMNT. These subroutine programs are hand coded IO programs to make them compatible with the FORTRAN IO system used by SEGMNT.

2. The present version of LOOK returns all data to the IWDLST array. The normal data = the concept numbers and syntactic indicators = are returned to the last five words of the IWDLST item, as previously explained. For each word not found in the dictionary, two items are returned: the location in the word of the first letter for which no match was found, and a code indicating whether this failure occurred while looking for a stem (1) or a suffix (2). These items are stored in the sixth word of the IWDLST item in the address and decrement, respectively. The sign of this sixth word is set to negative in order to distinguish it from ordinary concept numbers. These items are only stored if NOTFND is on; otherwise the word is

set to zero. In any event words seven to ten are set to zero if the word is not found.

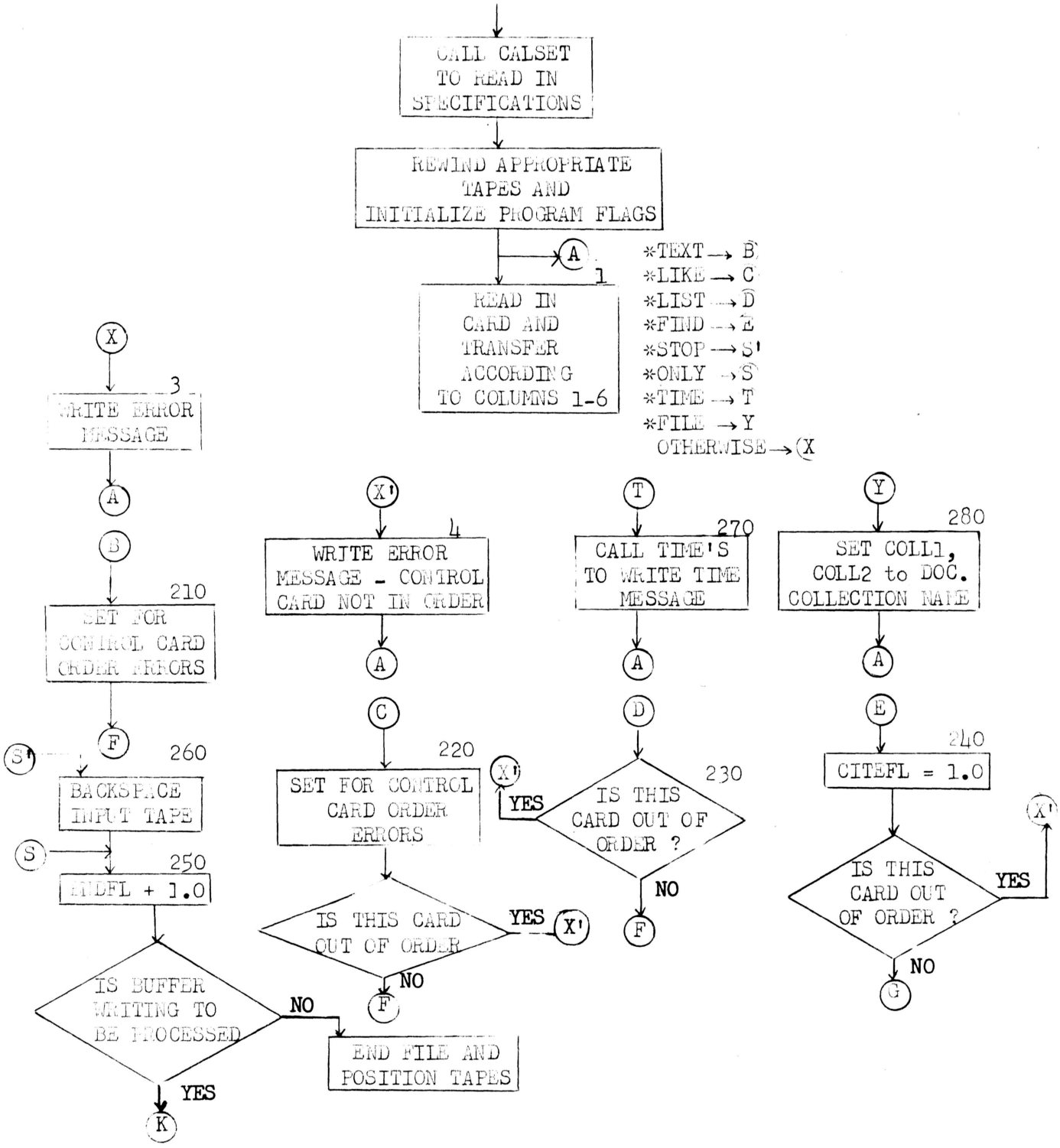
Subroutine NFRITE produces the tapes used by the second link to prepare the "words not found" output. For each document, a thirteen word record is written onto tape B1. The first twelve words are an image of the title card for the document; the decrement of the last word contains a count of the number of words not found for this document. The "words not found" items are written onto tape B2, each record containing only items from a single document. Each item written on tape consists of the first six words of the correponding IWDLST item; that is, the sentence and word-in-sentence numbers, the BCD word, and the not found codes left by LOOK. After processing a word, NFRITE resets the sixth word of the IWDLST item to zero.

If statistical or cluster phrases are to be produced, subroutine PHROCC is called. This program is described in detail by G. Shapiro in a later section of the present report. Subroutine VECFM is then called to convert the entries in IWDLST and the information left by PHROCC into a vector of the concepts that have been assigned to each document. The format of this tape is described in the introduction to this section. Each entry in the vector is a single word: The left half word containing the concept number, the tag field indicating its source, and the address field containing a weight which is a function of the frequency of occurrence of that concept in the document. For each occurrence of a test word associated with a given concept number the concept is given the weight 12/N, where N is the total number of concepts associated with that word. N therefore

varies from one to six, depending on how many of the six concept fields available in the dictionary format are used.

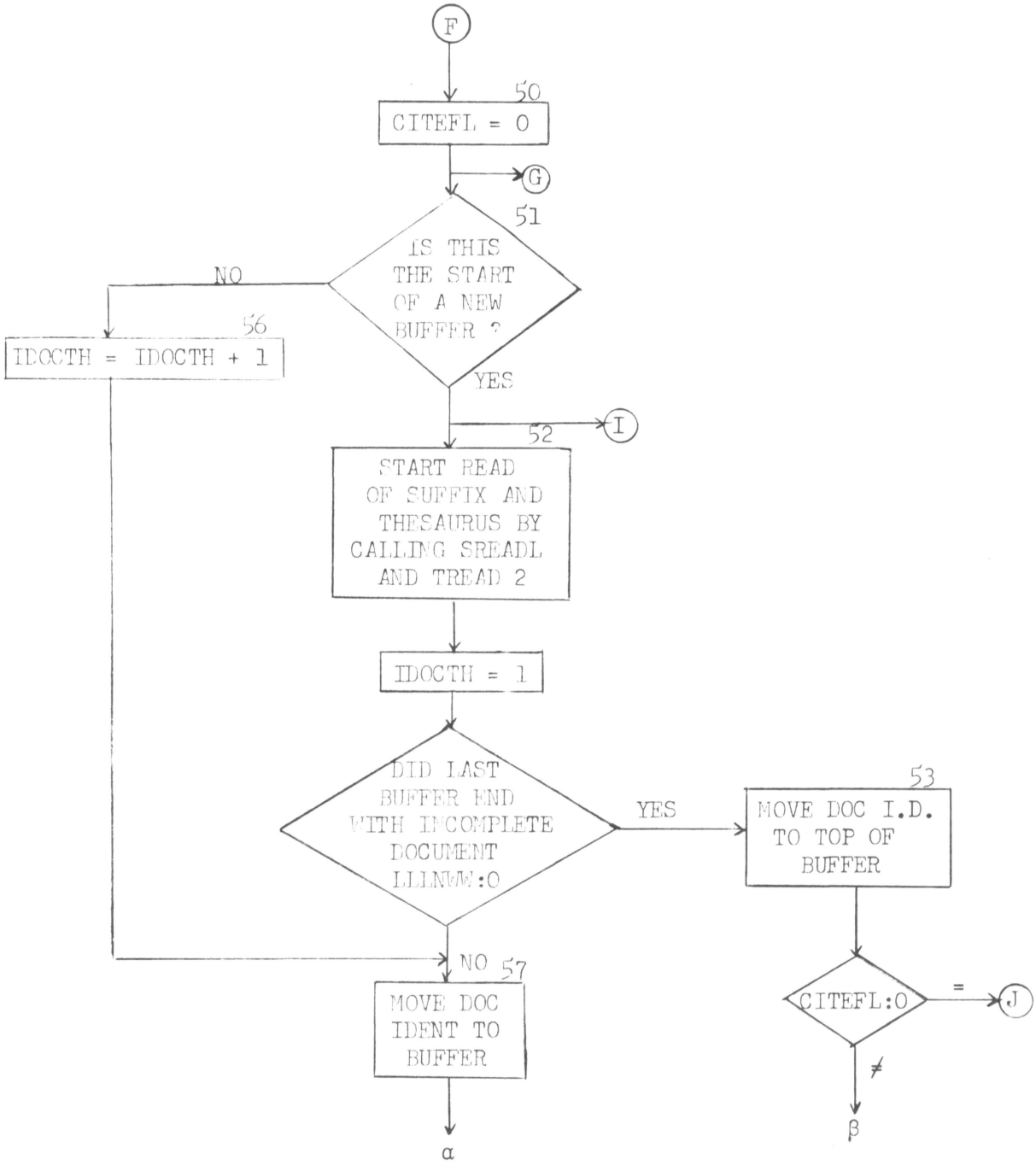
VECFM sets up a chained list of concepts for each document. When all words in a document have been processed, VECFM goes through the list and assembles the vector entries into a buffer which is then written onto tape. If only a part of a document fits into a buffer load, the partial chain of concepts is written onto tape. When the next buffer load is processed, the partial chain is read back and new entries are added to it.

The processing of a buffer load is completed when VECFM completes its work. The supervisor then checks the flag ENDFL to determine whether a *ONLY or a *STOP card has indicated the end of link one processing. If so, the supervisor positions tapes and calls ENDEND to go to the next link. If not, it checks to see if the buffer load just processed ended with an incomplete document. If not, it transfers to "A" on the flowchart to read in a new control card. Otherwise it transfers to "I". When SEGMNT is next called its own internal flags will indicate that the second part of an incomplete document is to be processed, and the corresponding items will be entered into the initial part of the IWDLST buffer. Flowchart 1 summarizes the complete for link one of the extended SMART system.

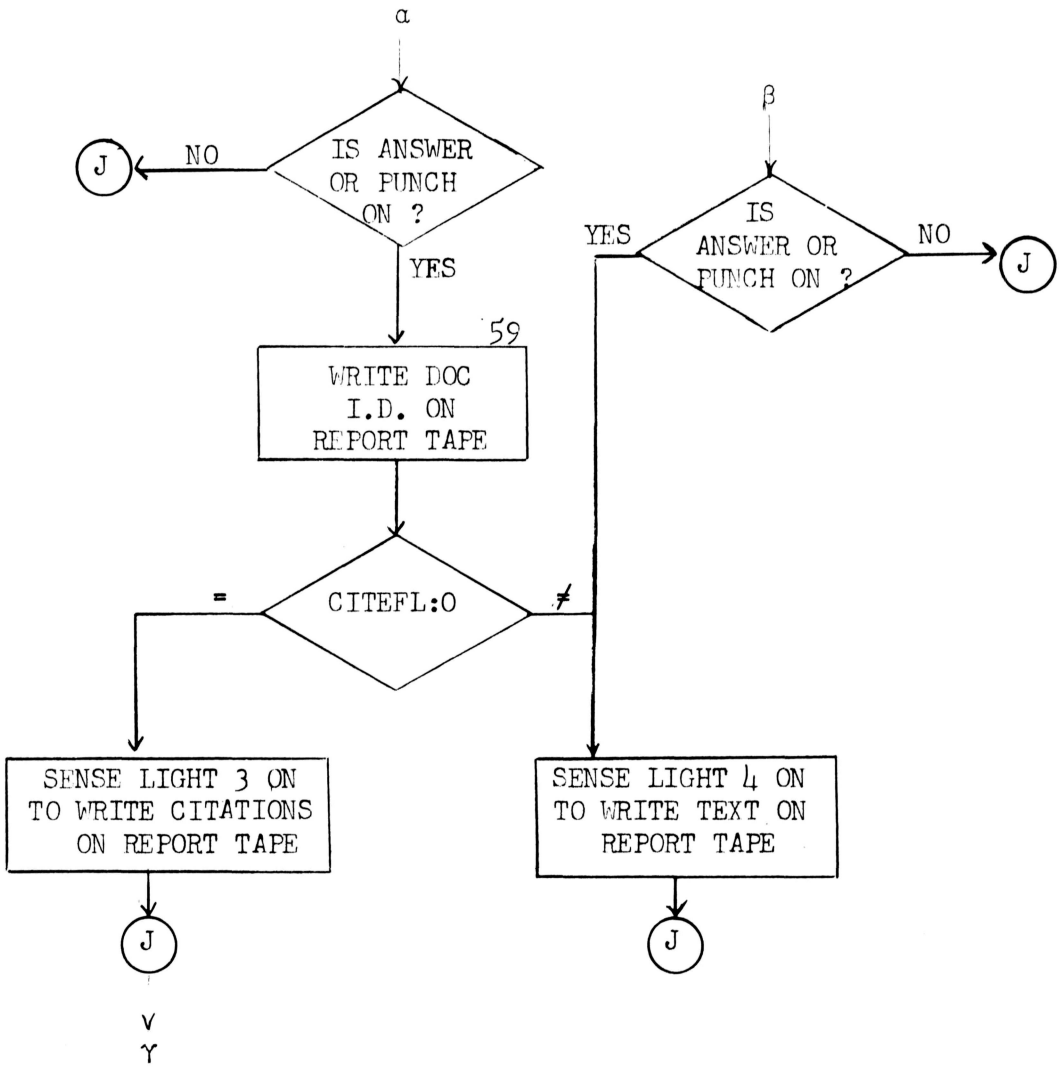


Super 1 – First Link Supervisor

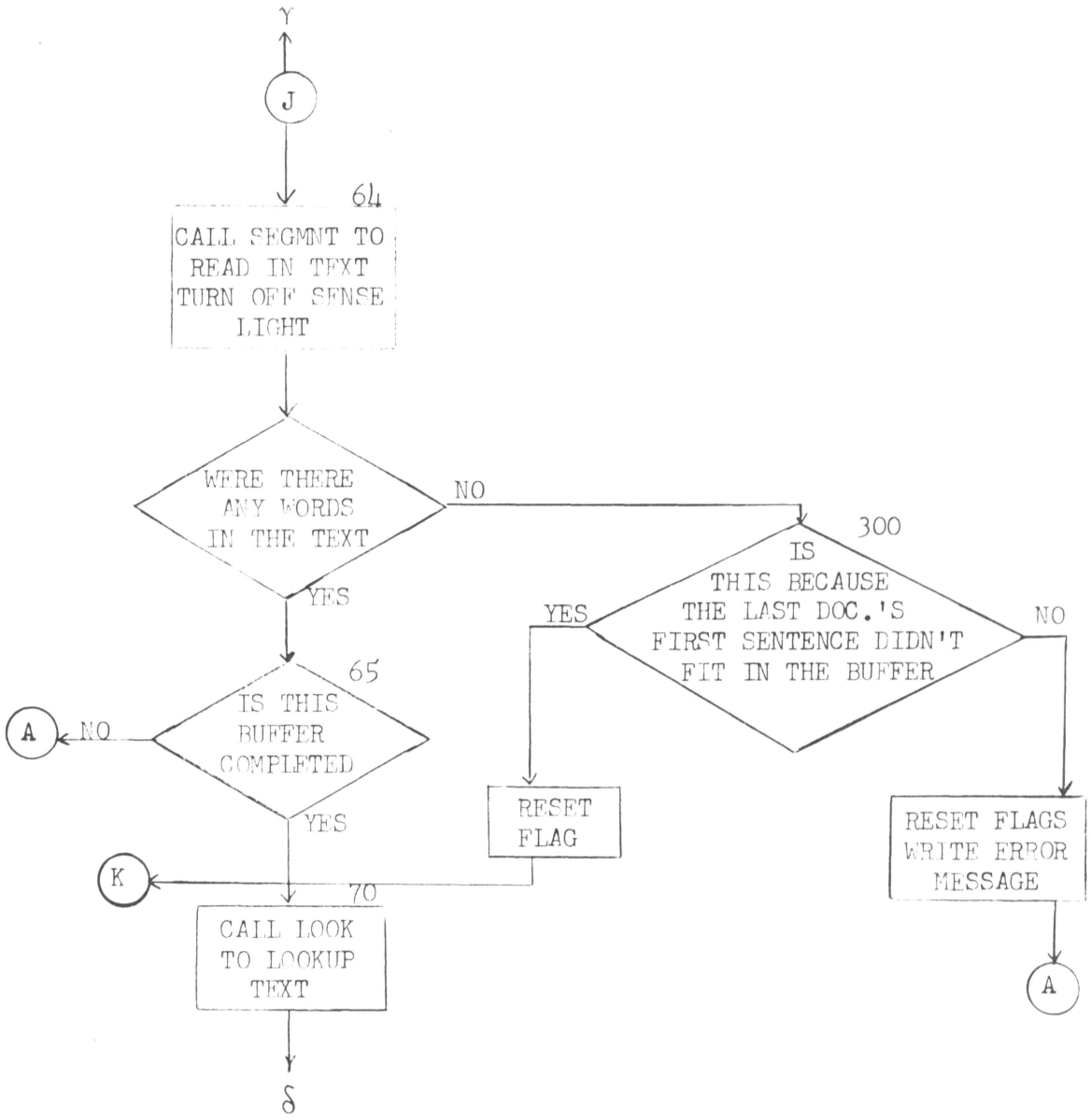
Flowchart 1



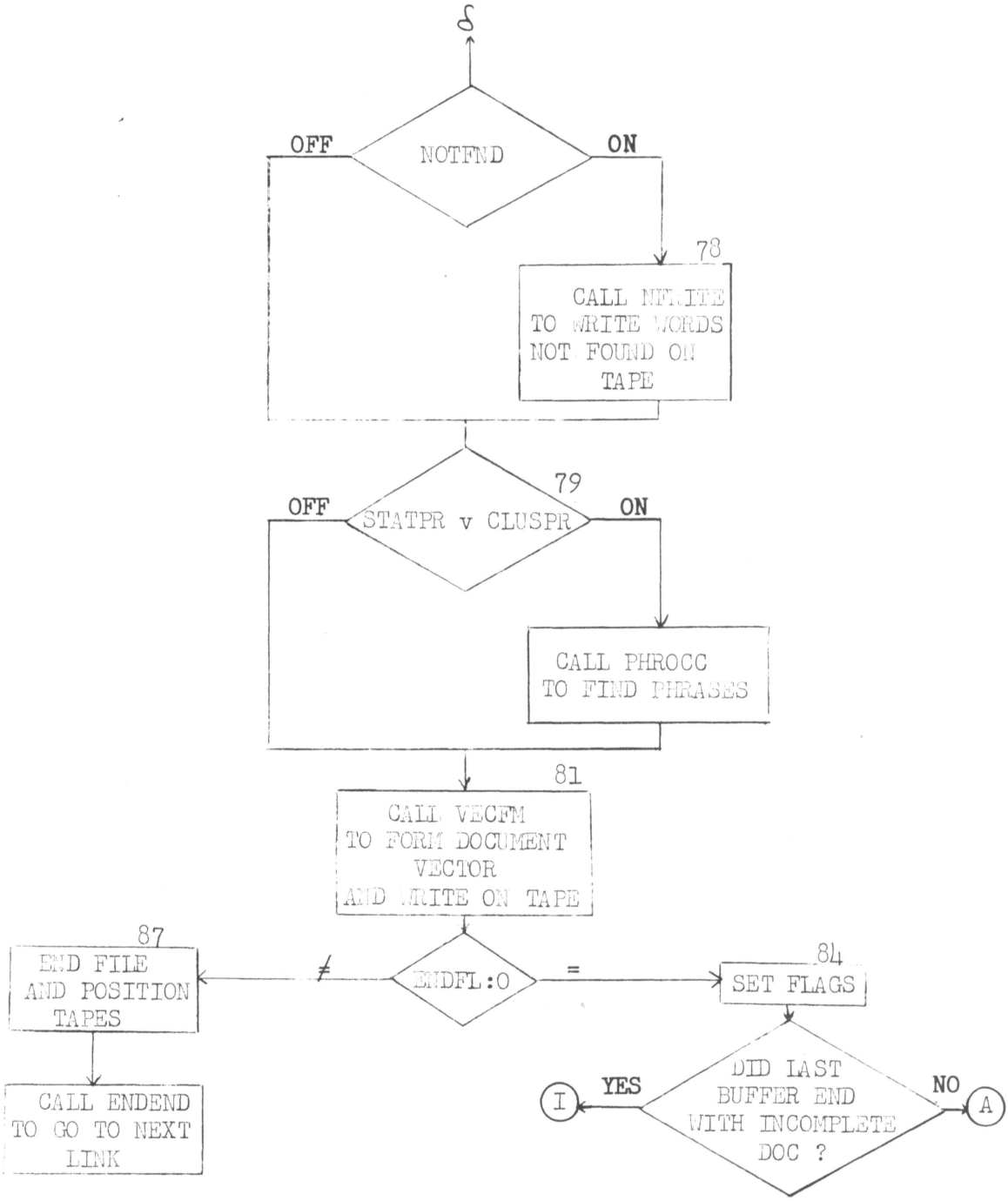
Flowchart 1 (continued)



Flowchart 1 (continued)



Flowchart 1 (continued)



Flowchart 1 (continued)