IX. STATISTICAL PHRASE PROCESSING

Michael Lesk and Tom Evslin

1. Introduction

A number of procedures have been included in the SMART system for the manipulation of selected sets of concept numbers. More specifically, it is desired to obtain frequency counts of the co-occurrences of these concept numbers within the sentences of a document. These concept sets are termed statistical phrases, by analogy with the structural (criterion) phrases in which the concepts are syntactically related.

The statistical phrases themselves are not initially derived from the texts, but rather are stored as file four of the library tape. Each statistical phrase is regarded as a semantic entity, and is provided with a concept number and a BCD name or "index." The present section deals with the routine used to construct and update this file, and with the routines that count the occurrences of the corresponding phrases within the document collection.

2. Updating of the Statistical Phrase File

CRITS2 writes the fourth file of the library tape, consisting of a dictionary of semantic data about phrases. This file provides part of the information needed by the criterion tree routine. It is also used by the statistical phrase searcher, PHROCC, to find phrases in texts. For each phrase, this file includes:

(1) a six-character BCD "identifier" or "index";

(2) the concept number of the whole phrase, called the "output concept number";

(3) up to six "component concept numbers," the co-occurrence of which in a sentence is counted as a phrase by the statistical phrase searcher.

This information is placed in a four machine-word item. The first word is the identifier; the second contains the output concept number in the decrement. Each of the last two words holds three twelve-bit component concept numbers; zeros are used for fill. For example, "parity" is represented by concept number 207, and "check" by 271; the phrase "parity check," however, corresponds to concept 300. The entry for this phrase is shown below:

WORD 1	2	3	4
PARCHK	300	207, 271, 0	0, 0, 0
Identifier	Output Concept Number	Component Concept Numbers	

(The numbers, of course, are converted to binary internally.)

To improve efficiency, the items are blocked 100 to a physical record on tape, the last record being possibly short. A three-word label is prefixed to each record, thus resulting in a length of 403 words for all but the last record in the file. The last record is $4k+3$ words long, where $1 \leq k \leq 100$. The file is stored in ascending order by phrase identifier (index).

Subroutine CRITS2 operates by a straightforward merge pass of the old library file on A6 with a set of correction cards on A2. The first card on A2 is an update control card, containing one of four code words in columns 1-6. The options are:

MERGE	Merge correction cards on A2 with the file on A6.
COPY	Copy old file from A6 to B5; no correction cards follow.
REPLAC	Replace (the "E" may be punched in column 7 if desired) the old file on A6 with the file that follows on A2, spacing over the old file on A6.
IGNORE	Ignoring tape A6 completely, write a new file on B5 from the cards following on A2. This option is used for new starts only, since in a tape update run tape A6 would be left mispositioned.

The control card may contain either PRINT or PUNCH in columns 7-11, producing a copy of the new file on either the print or punch tapes, respectively. Random characters should not be punched in column 9 of this card, because this may be interpreted as one of the permissible specifications.

If any option other than COPY is specified, the control card must be followed by a set of correction cards for the file. These are in the form of entries to the file, and are punched in the following format:

(1) six blanks in columns 1-6;

(2) the phrase identifier, six BCD characters, in columns 7-12;

(3) the output concept number in columns 17-20, right-justified;

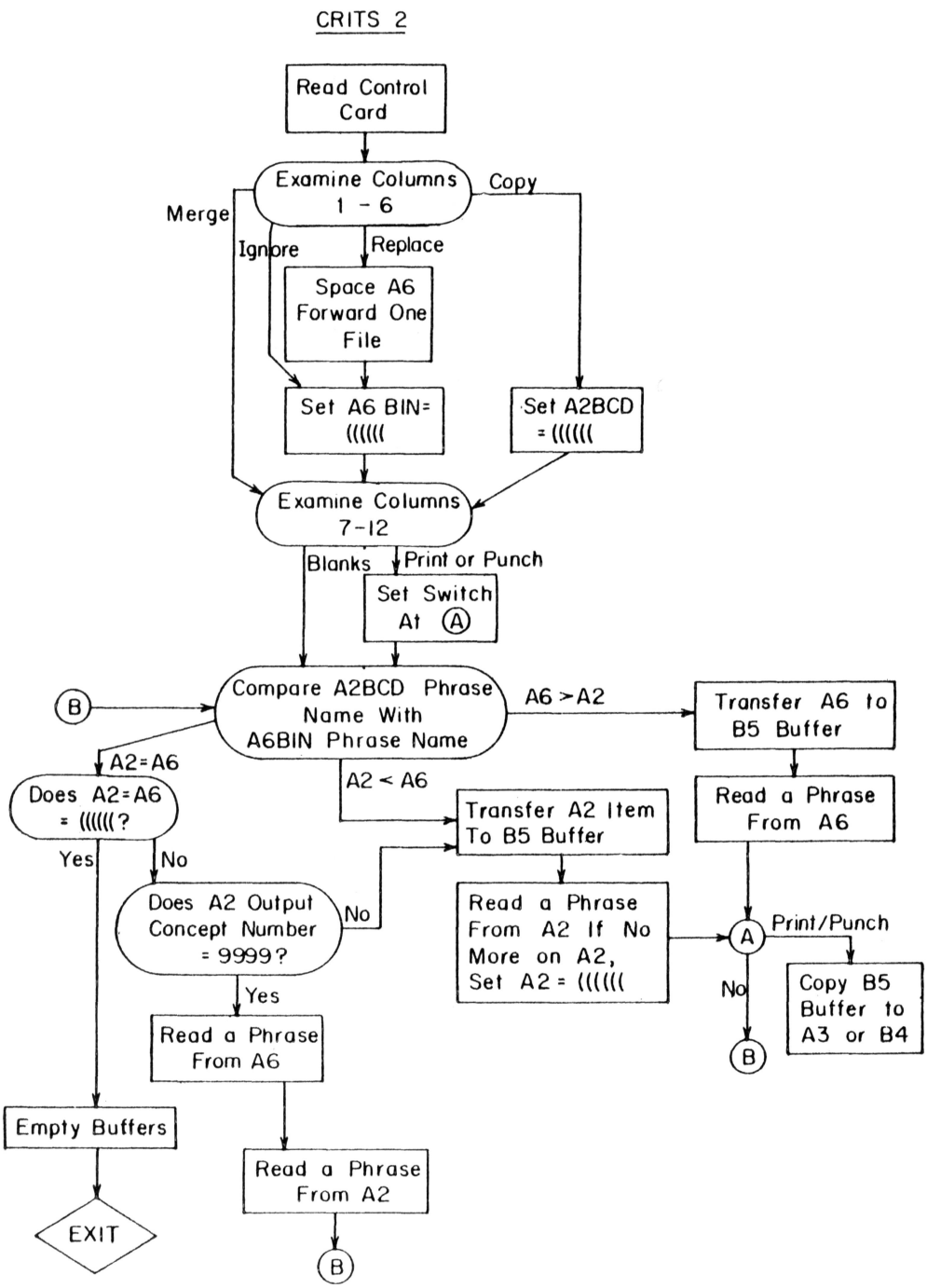
(4) up to six component concept numbers in columns 21-50, consisting of five columns each; within each five-column entry, the number must be right-justified.

The control cards should be in ascending order on the phrase name. No special ending sentinel is needed. The phrase identifier must not consist of six left parentheses, and in the interests of readability should not be six blanks.

In the MERGE option, if an identifier on A2 does not match an identifier on A6 it will be inserted in the proper place. If it does match an A6 entry it will replace the old item, unless the A2 card has an output concept number of 9999. In this case, the phrase is deleted from the file.

CRITS2 makes use of a set of FAP subroutines to perform tape read/writes. The subroutines call the FMS input-output package, the first record of the system tape. The entry points to these subroutines are:

BACKA2	Backspaces tape A2 one record.
IOP	Loads system input-output package (IOP) from A1.
READA6	Delivers one item from A6 to input area A6BIN, unblocking as needed. When the current record is completely processed, the next block is read from A6. If all phrases in the old file have been read, a phrase name of six left parentheses is returned, indicating end of file.
WRITB5	Moves one item from output area B5BIN to output buffer. When 100 items have accumulated, a record is written on B5.



Statistical Phrase Update Procedures

Flowchart 1

COMPAR Compares identifiers in input areas A6BIN and A2BCD. Control is returned to the first, second, or third location after the calling instruction if the A2 identifier is respectively less than, equal to, or greater than the A6 identifier.

GREATR Compares its first and second arguments and sets accumulator to zero if the first is greater than the second; nonzero otherwise.

TCOB or WEFB (equivalent) writes out any material waiting in B5 buffer and writes end of file on B5.

SPACE6 Forwards tape A6 one file.

Note that on the 7094 six left parentheses, corresponding to the octal number 747474747474_8, constitute the highest-sorting combination of six BCD characters.

A flowchart of the CRITS2 routine appears as Flowchart 1.

3. Statistical Phrase Counting

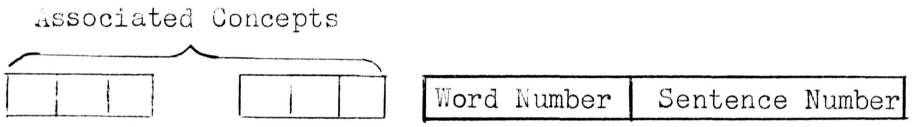
PHROCC is a FORTRAN subroutine to count the occurrences of statistical phrases in a given document. EVSINP is a short FAP subroutine called by PHROCC to provide binary input.

The phrases appear as records of 403 words each on B5, terminated by an end of file. The first three words of each record are not processed. The remainder of the record is made up of 100 or less four-word items as shown in Part 2 of this section.

EVSINP reads one record at a time into an array whose FORTRAN address is given by the second parameter in the calling sequence. If less

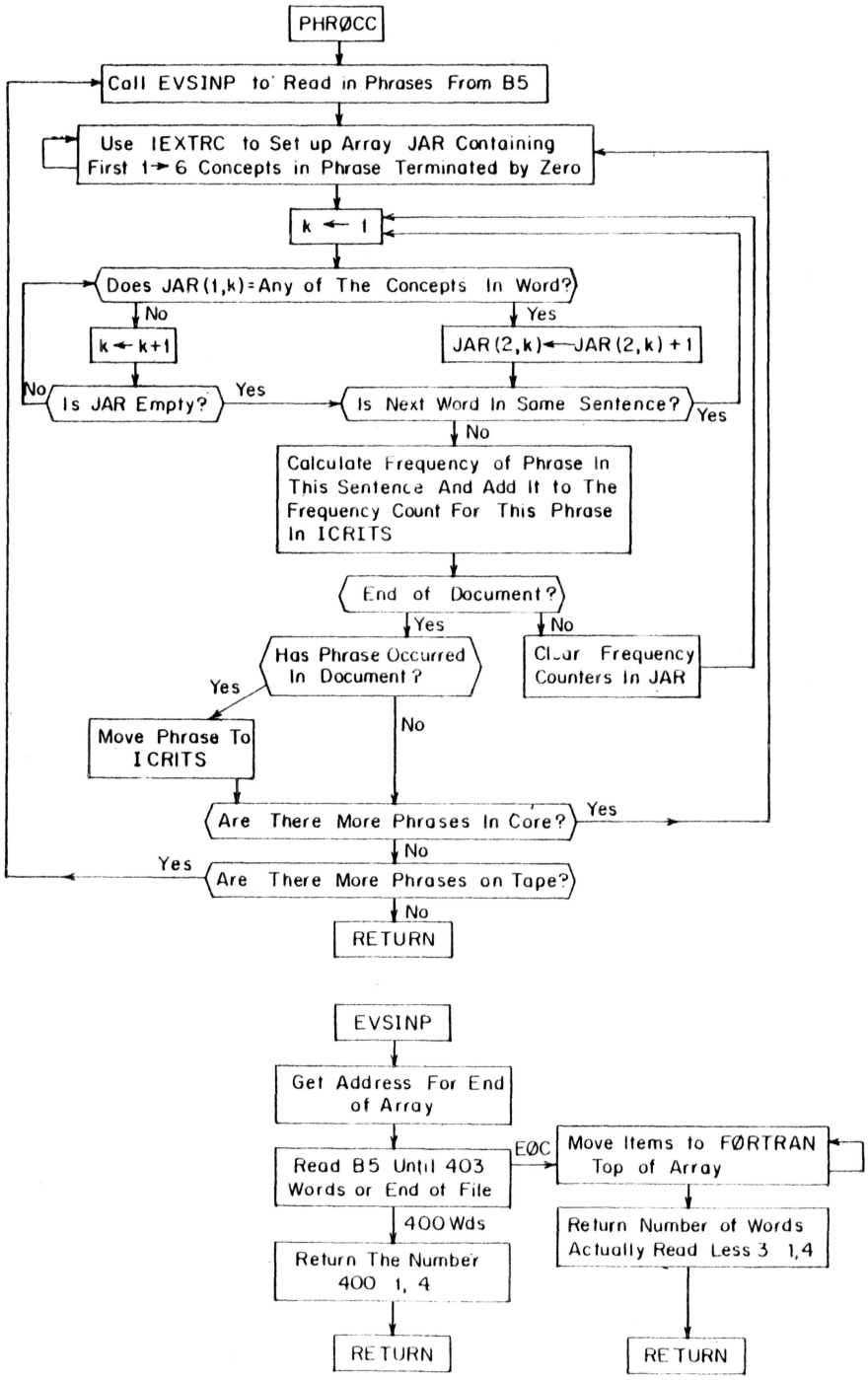
than 100 four-word phrases were found in the record, EVSINP repositions the phrases in core so that they will be in the proper place for the indicated FORTRAN array. The number of words in the record, excluding the three junk words at the beginning, is returned to the location indicated by the first parameter in the calling sequence. This parameter is made negative to indicate that an EØF was found while reading the current record, and that no further phrases are present on tape.

PHRCCC takes each phrase in its buffer and compares its component concepts to the concept numbers associated with the words of the document stored in WDLST. Each word in the document consists of three computer words as follows:



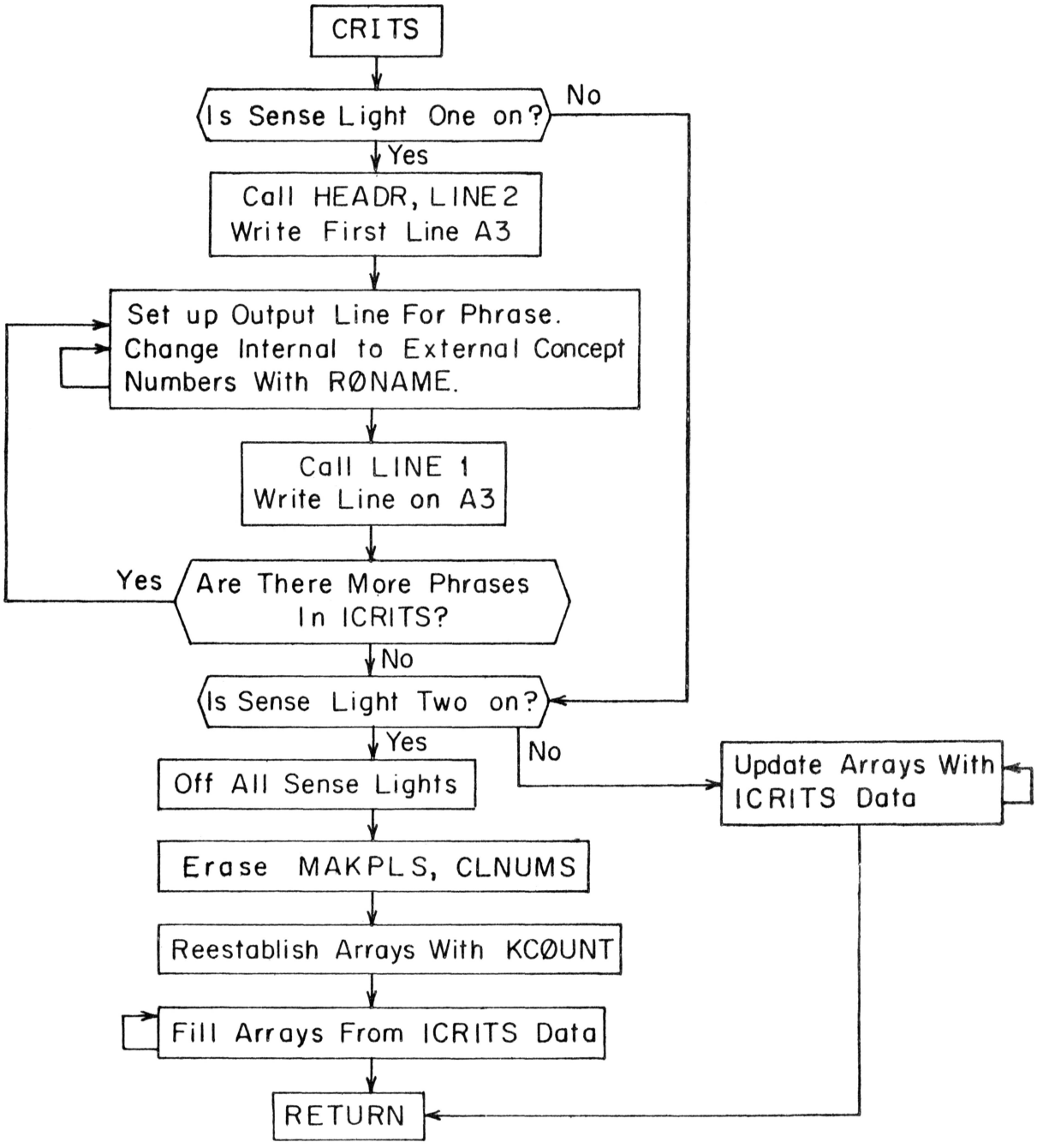
Each concept number of the phrase is compared to the concept numbers of each sentence word, and matches are counted. Whether or not a match is found, all concepts of the phrase are compared to the concept numbers of the next sentence word iteratively until the sentence is exhausted. The frequency of occurrence of a given phrase in a given sentence is defined as the number of matches made by the least matched concept of the phrase. This process is repeated sentence by sentence until the entire text has been scanned.

The occurrence counts of each phrase in each sentence are added together to give the total frequency for the particular phrase in the current



Statistical Phrase Counting

Flowchart 2



Output of Statistical Phrase Count

Flowchart 3

document. If a given phrase occurs at all, it will be entered into the FORTRAN array ICRITS along with its frequency count.

After all the phrases on the tape are processed, PHROCC returns control to CHIEF. A flowchart appears as Flowchart 2.

4. Output of Phrase Counts

CRITS is a FORTRAN subroutine which prints the list of phrases detected and adds the frequency counts to the document's concept-vector.

Printing is controlled by sense light 1. If it is on, the phrases stored in array ICRIB are listed in the following format: BCD identifier, output concept identifier, component concept identifier.

Addition of phrase data to the concept-vector of the document can be done in two ways if statistical clustering was used. The new phrase data may add to or supersede the old statistical data. If sense light 2 is on, the new data replaces the old. The statistical clusters are "forgotten" and the concepts recounted by subroutine KCOUNT. If sense light 2 is off, the new phrases are added to the existing counts in KLSOCC, the array of concept occurrences, without first erasing the clustering data. The output procedure is shown in Flowchart 3.