ISR-9
August 1965

III. A SPECIFIC DESCRIPTION OF THE NEW SMART SYSTEM

T. Evslin

1. Operating Instructions

A. Input Parameters

The first data cards introduced into SMART at object time are the option cards containing the user-specified input parameters. The various options listed below must be punched according to the following format: specifications may appear in columns 1-80 of the cards; there must be no blanks within specification names, and no punctuation must appear on the cards except the decimal point in floating point numbers; blanks must appear between parameter names; any number of blanks are treated as a single blank. Specifications requiring a second word (see ANSWER below), or a number (see CUTRD) are followed by a blank, and then by the word or number. Such specifications are marked below with an asterisk. Any number of cards may be used. The last card is terminated by the character X. Parameters not set by the user will be set to an initial value listed below.

Five types of input parameters are used: those calling for printouts, those specifying logical operations to be performed, those setting system constants, those specifying how an option should be performed, and those identifying the run. Where classification of a given parameter is dubious, parameters have been listed under all possible classifications.

(a) Print Options

All print options are initially set to zero to indicate that the output in question will not be performed unless otherwise specified. If the name of the parameter appears in the control deck, the parameter is then set to a nonzero value and the corresponding output task is performed. The parameter is set to one unless it is identified by a second name, in which case the possible values are discussed below.

Name	Function
PRNVEC	print vectors
NOTFND	print words not found in lookup
NODECO	print node correspondences
SYNANA	print syntactic analysis
PREQCO	print request-document correlations
PDOCCO	print document-document correlations
PCOCOR	print concept-concept correlations
PCONRD	print concordance
ENGTXT	print English text during lookup
ANSWER*	print answers to requests (second parameter must follow: SHORT sets ANSWER to one prints two-word identifiers only MEDIUM sets ANSWER to two prints titles only LONG sets ANSWER to three prints text of requests and citations of documents if ANSWER= 0 and SCORES= 0, no request-document correlations are performed)

(b) Logical Options

The parameters for logical operations are given an initial value of zero. Those which are not specified by a second word are set to one when names by the user. Others are set as specified. A value of zero implies that the operation is not performed.

Name	Function
ANSWER*	perform request-document correlation (see ANSWER above)
SCORES	perform evaluation, perform request-document correlation
STATPR	find statistical phrases
CLUSPR	find cluster phrases
SYNTAX	find syntactic phrases
PUNCH	punch out looked-up texts
CONCON*	perform concept-concept expansion (CONCON must be followed by an integer specifying the number of times concept-concept correlation should be iterated)
DOCDOC	perform document-document correlation
HIER*	perform hierarchical option (HIER must be followed by an additional parameter EXPAND sets HIER to one to signify a hierarchy expansion SHRINK sets HIER to two to signify a hierarchy replacing)
DOCTAP	use documents from permanent collection tape

(c) System Constants

Parameters specifying system constants are given differing initial values as listed below. If the user wishes to set the value of one of these constants, he must specify the parameter name followed by a floating-point number to set the constant.

Name	Function	Initial Value
ROOTWT*	weight of parents	0.0
BRANWT*	weight of brothers	0.0
LEAFWT*	weight of sons	0.0
CROSWT*	weight of cross references	0.0
COCOWT*	weight of concept-concept expansions	0.0
STATWT*	weight of statistical phrases	0.0
SYNWT*	weight of syntactic phrases	0.0
CLSWT*	weight of cluster phrase	0.0
STEMWT*	weight of word stems	1.0
TITLWT*	weight of words occurring in title	1.0
BODYWT*	weight of words occurring in body	1.0
CUTRD*	cutoff for request-document correlation	0.35
CUTDD*	cutoff for document-document correlation	0.50
CUTCC*	cutoff for concept-concept correlation	0.90

(d) Operating Mode

These options have differing initial values as listed below. Most of them must be followed by a second parameter. The corresponding options are, as usual, marked by an asterisk.

Name	Function	Initial Value
EXPAND*	apply expansion options (CONCON and HIER) as follows: REQS sets EXPAND to one, requests only DOCS sets EXPAND to two, documents only ALL sets EXPAND to three, both	1
LOGVEC	sets all weights to one if calles (each concept is given equal weight regardless of frequency of occurrence)	0
MODEDD*	mode of document-document correlation: COS sets MODEDD to one, use cosine correlation OVLAP sets MODEDD to two, use overlap correlation	1
MODERD*	mode of request document correlation (see MODEDD)	1
MODECC*	mode of concept-concept correlation (see MODEDD)	1

(e) Run Identification

Name	Function	Initial Value
THES*	must be followed by signed integer giving the thesaurus number to be used (+ handmade; - null)	+1
MAXCON*	must be followed by an integer giving the largest concept number in dictionary	32000

B. Control Cards

All control cards for the SMART system must have an asterisk in column 1, the four-letter control word in columns 2-5, and a blank in column 6. The rest of the card is not looked at except in the case of *FIND, *LIST, *LIKE, and *TEXT cards where the document title is given, and in the case of *FILE cards.

If a collection tape is being used and it contains more than one collection, a *FILE card with either a BCD identifier or an integer in columns 7-12 may be used. An identifier is assumed to be the name of the collection to be used; a number indicates the position of the collection on the tape.

*TIME causes the current time and the time elapsed since the beginning of the execution to be printed out off-line.

*ONLY must precede the first binary deck of a text looked up during previous runs. The *ONLY card must be used even if there are no texts to be looked up in this run.

*STOP terminates the deck of documents not looked up, or previously looked up, which are introduced during the current run. The *STOP card must appear before the prediction decks whether or not there have been any such texts.

If any of the *FIND, *LIST, *LIKE, or *TEXT cards appear before the *ONLY card, the text following them is looked up. If these cards appear after the *ONLY card, they must precede the binary deck of a document previously looked up. *FIND and *LIST cards both identify a request. The only difference is that citations are ignored following *LIST cards. *LIKE precedes a document which is also to be treated as a request and correlated with all other documents. *TEXT indicates a "pure" document which will be correlated with requests only.

C. English text Following *FIND, *LIKE, *LIST, and *TEXT Cards

The remainder of the card following the control word contains the document title in columns 7-72. The cards immediately after the title card, if any, contain citation data and begin with a "$" in column 1. They may be punched out to column 72. The body of the text is punched in English in columns 1-72, following the citations. A word may be split at the end of a card if the last character on the card is a hypen. The first character of the remaining portion of the word must appear in column 1 of the next card.

A quotation mark is replaced by a slash. A semicolon becomes ",.". A colon is replaced by "..". Question marks are represented by ".QUE". A period at the end of a sentence is preceded by a blank.

D. Evaluation Prediction Decks

The first card of the prediction deck contains a three-digit number, prefaced if necessary in columns 1-3. This specifies the number of requests submitted during this run, as well as the number of prediction decks which follow.

Each prediction deck begins with a card containing a two-word request identifier (the first two words of the request name) in columns 1-12, and the number of relevant documents in columns 13-16(the number of prediction cards to follow for this request).

The prediction cards follow the request identifying card in descending order of predicted correlation with the request. Each prediction card carries a two-word document identifier in columns 1-12.

E. Order of Input Deck

The first cards following the *DATA cards must be used for input parameter specification. Following these the documents to be looked up during this run must be introduced, preceded by *FIND, *LIST, *LIKE, and *TEXT cards. Documents preceded by *FIND and *LIST must be submitted first, followed by *LIKE, and last by *TEXT cards.

The *ONLY card precedes the binary decks of documents to be looked up. These decks will again be headed by *FIND, *LIST, *LIKE, and *TEXT cards in the same order as above.

The *STOP card precedes the list of predictions which forms the last deck before the end-of-file marker.

The *FILE card, if any, may come anywhere before the *ONLY card. *TIME cards may precede any of the control cards.

F. A Sample Input Deck

```
1  7     12

* DATA
PRNVEC ENGTXT CONCON 2 EXPAND 1
DOCDOC SCORES DOCTAP CUTCC 0.86
STATPR STATWT 2.0 TITLWT 1.0
COCOWT 3.0 TITLWT 1.5 MODECC
OVLAP THES +3 MAXCON 10000 X
*TIME
*FIND CRYSTAL CALCULATION
CRYSTAL CALCULATIONS ARE
DIFFICULT.  PLEASE GIVE ME INFOR-
MATION ABOUT STRUCTURE FACTOR
CALCULATIONS
```

```
1     7    12
*FIND (identifier)
  (text)
*LIST (identifier)
  (text)
*LIKE (identifier)
$G..SALTON, INFORMATION RETRIEVAL
$ AND TEXT INDEXING
  (text)
* TEXT (identifier)
  (citations)
   (text)
*TEXT (identifier)
  (text)
*TIME
*TAPE 6
*ONLY
*FIND (identifier)
  (looked up request in binary)
*LIKE (identifier)
  (looked up like)
*LIKE (identifier)
  (looked up like)
*TEXT (identifier)
  (looked up text)
*TIME
*STOP
002
CRYSTAL CALC 2
CRYSTAL FACE
CRYSTAL STRU
INFORMATION 1
RETRIEVAL OF
eof
```

2. System Description

The following is a technical operating description of the SMART system. For a description of the system's logical functions and purposes see Sec. II of this report. For a detailed explanation of specific programs, see the section covering the appropriate chain link. A flowchart describing the complete programs follows later in this section. Tape allocations are explained in detail in Part 4 of this section.

The SMART system runs on the IBM 7094 under the FORTRAN II, version II, monitor system, with a nonstandard loader. This monitor system is available with SMART. Fourteen tape drives, eight on channel A and six on channel B, are needed to run SMART with all options working and at full efficiency. However, fewer tapes may be used as explained in Part 4 of this section.

SMART processes English language documents of any text length. However, these documents may not give rise to more than 32,000 distinct noncommon concept numbers, since this constitutes the thesaurus limit. An additional 32,000 concept numbers each are reserved for cluster, phrase, citation, author, and journal concepts; 767 concept numbers are reserved for common words. As many as 262,143 documents may be included in a collection. A maximum of 49 requests may be introduced at a time, although the existence of the merged correlation tape makes it possible to correlate an unlimited number of requests against any collection.

Link 1 of the SMART system (see Sec. V) is always called. The first task performed in this link is to read in the object time parameters from tape A2 and to set system constants. A list of specifications is written on tape A3, and all parameters are set in upper common.

If any text needs to be looked up during this run, it is read in from tape A2 one buffer load at a time. The thesaurus is read in from B5 for each batch. If phrase or cluster detection has been specified, the appropriate dictionary is read in from B5 and this operation is also performed on one batch load of text at a time (see Sec. VII).

English text is written out for printing on A3 if called for. Partial vectors including concepts derived from word stems, phrases, and clusters are written out on A5. Document titles and citations are written on A4. If there is to be a printout of words not found, the necessary information goes onto B1 and B2. If syntactic phrase processing has been specified, sentences containing statistical phrases are written onto A7.

Link 1 is terminated by the appearance of an *ONLY card on A2. If words not found are to be printed, link 2 is called next. If not, link 3 is called for syntactic processing; or, if no syntactic processing is to take place, link 8 directly follows link 1.

Links 3-7 comprise the Kuno syntactic analyzer slightly modified to run with SMART (see Sec. X). Link 7 is the tree matcher. Here phrase subtrees are matched against the syntactically produced sentence trees, and the concept numbers corresponding to syntactic phrases are written on A7.

Link 8 (see Sec. XIII) merges, weights, and correlates vectors. Partial vectors for looked-up requests and documents are read in from A5 and also from A7 if syntactic phrase searching has previously been performed. If the user has specified output punching, the looked-up documents are put out for this purpose on B4. A listing of merged vectors may be written on A3. At this time, previously looked-up documents from A2 and from A6 (the latter being the collection tape), are read in and weighted.

If there is to be immediate request-document correlation, requests are stacked in core memory and each document is correlated against these requests as it is passed through memory. However, if concept-concept or hierarchy expansion must take place, the vectors are written out onto B1 and are neither correlated nor stacked. If there is to be document-document correlation, the documents only are written on B1 unless this tape contains the results of an expansion, in which case the documents go out on B6. Correlations are written onto B2 unless this tape is tied up by an expansion, in which case they go onto B6.

Link 9 (see Sec. XIV) and link 10 (see Sec. XV) perform, respectively, the concept-concept and the hierarchy expansions. In link 9 the document-document matrix is inverted and each concept is correlated with each other concept. This process may be iterated several times as specified by the user.

In link 10, significantly related concepts derived from the concept-concept expansion and hierarchically related concepts are added to the document vectors. The matrix is reinverted and the output appears on tapes A5, A7, A8, B1, B2, or B6, depending on the output tape chosen by the high speed sort. Control is returned to link 8 which merges and weights the expanded vectors and performs request document correlation as previously explained.

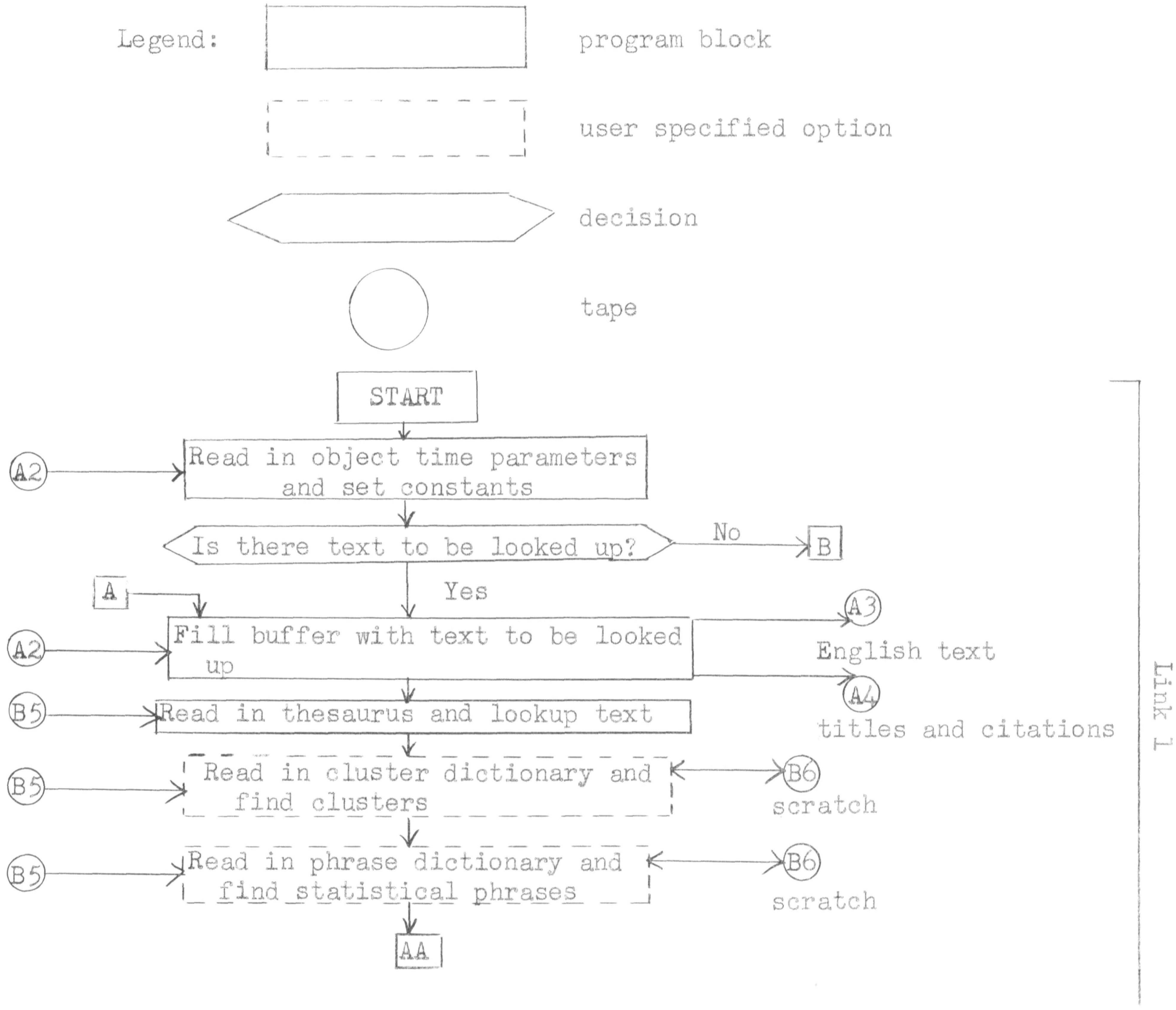
Document-document correlations, if any, are performed in link 11. Each document is correlated against each other document. The results are used to expand the list of documents returned as answers to each request. For a fuller description of this process, see Sec. XIV.

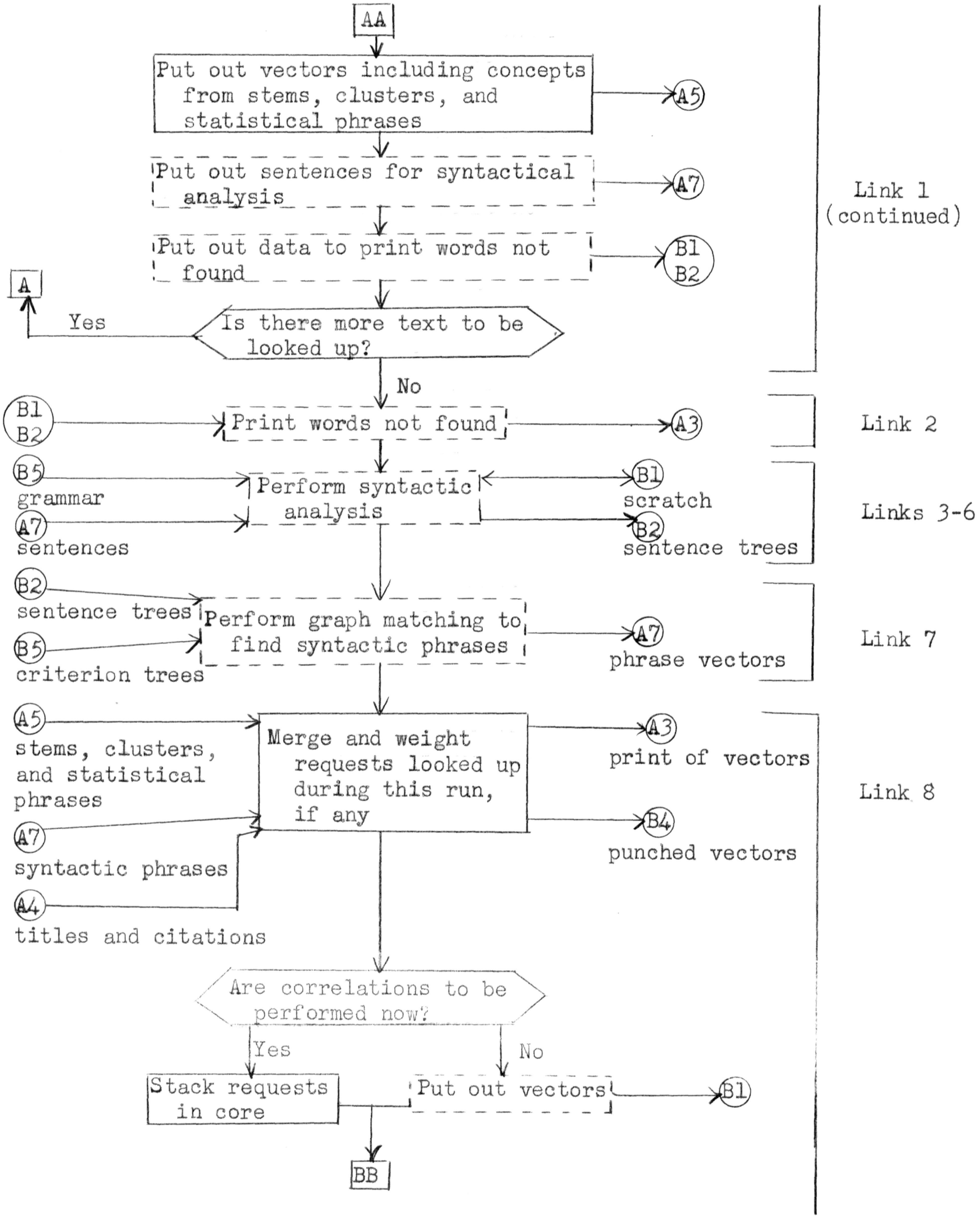
If the user has asked for a printout of requests, answers, and correlations, this is produced by link 13 following a sort of the correlation tape in link 12. Three types of output are available to produce varying amounts of document output (see Sec. XVI).

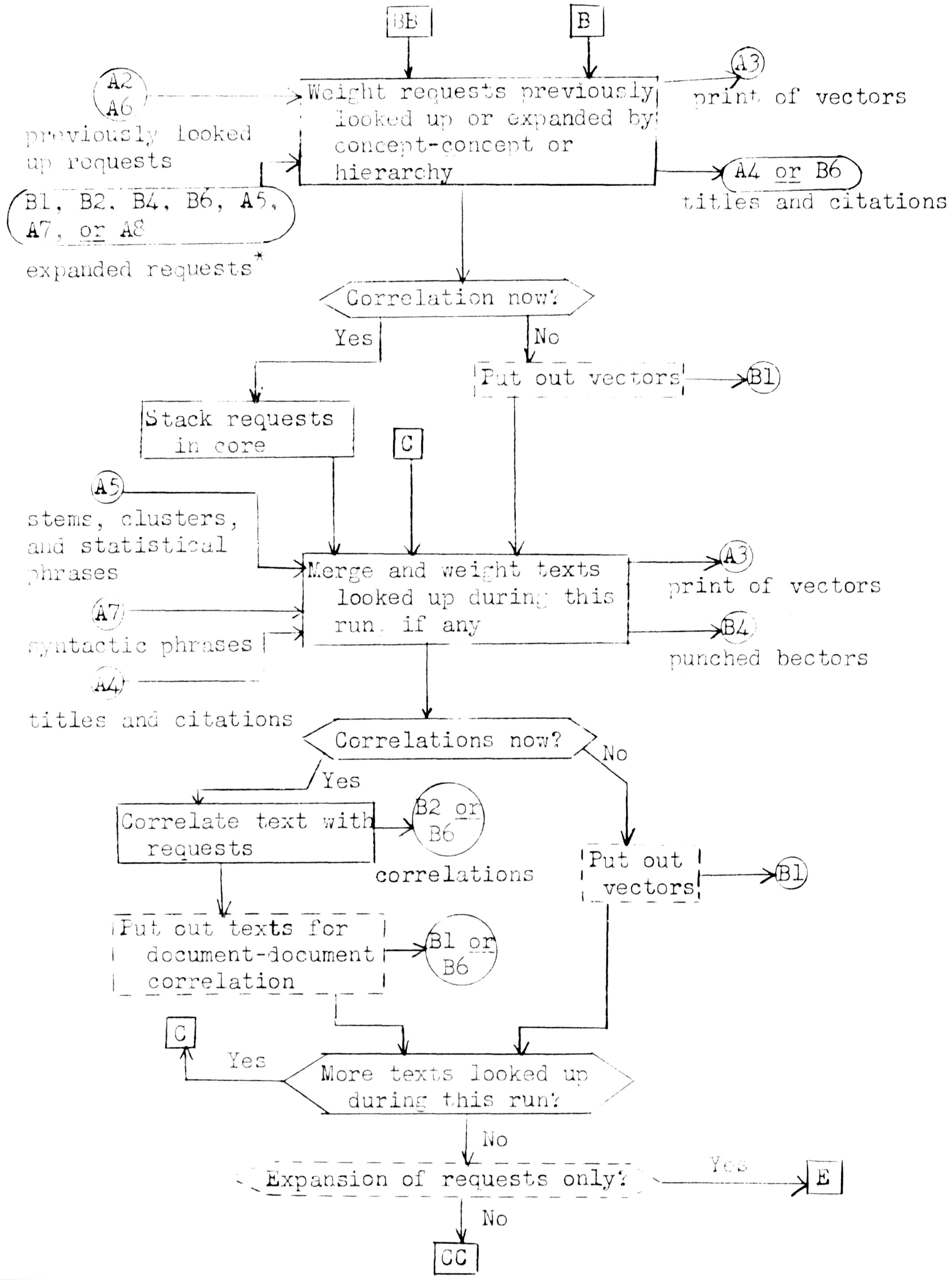
Evaluation of answers and correlations is done in link 14 (Sec. XVII). Predictions are read in from tape A2 and matched against actual answers. Evaluation output is printed on tape A3.

A post processor is available for the system which merges the correlation tape and the run parameters into a merged result tape to be used later for further evaluation.

3. System Flowchart

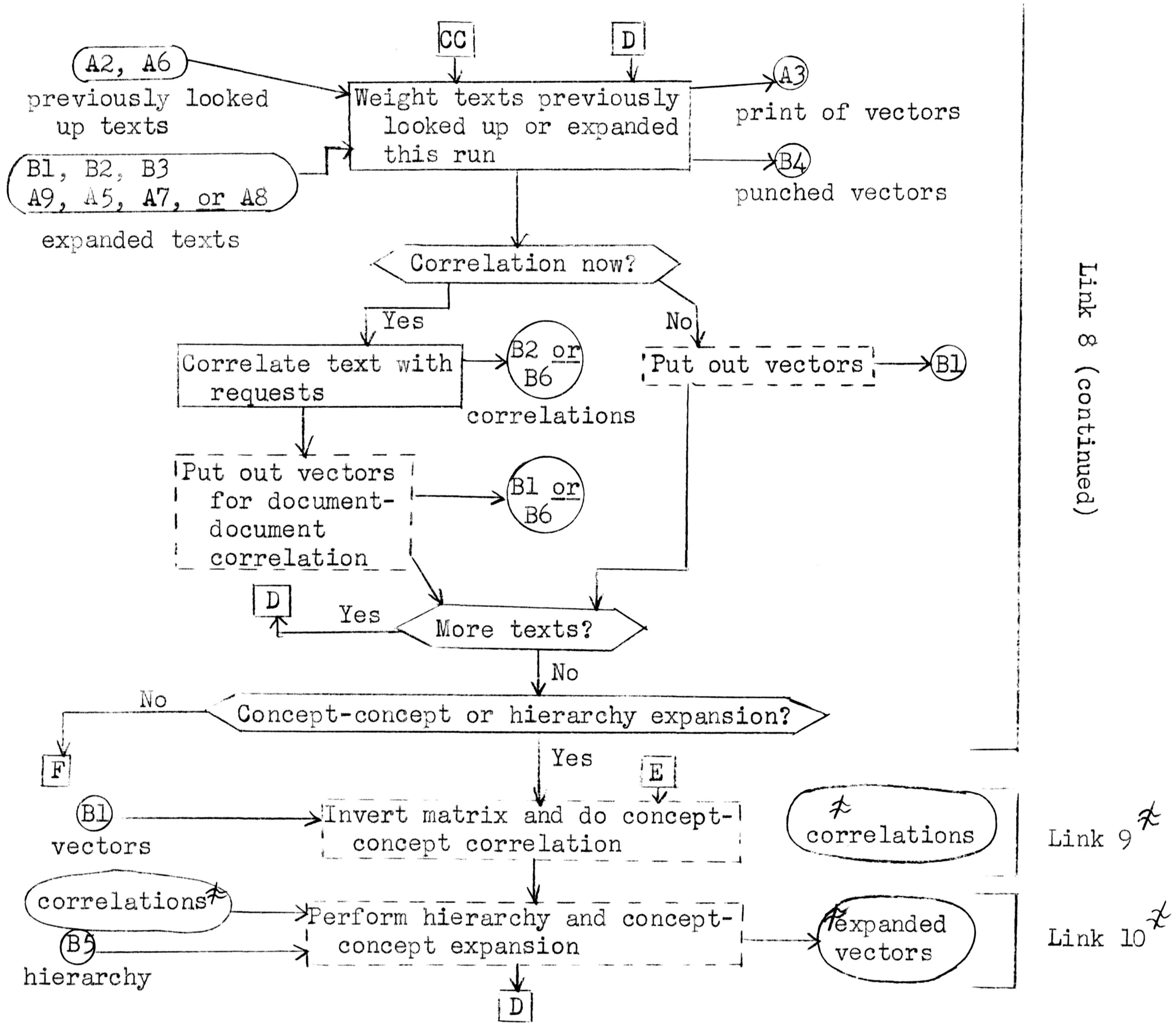

Legend:
program block
user specified option
decision
tape
START
A2
Read in object time parameters and set constants
Is there text to be looked up?
No
B
Yes
A
A2
Fill buffer with text to be looked up
A3
English text
A4
titles and citations
B5
Read in thesaurus and lookup text
B5
Read in cluster dictionary and find clusters
B6
scratch
B5
Read in phrase dictionary and find statistical phrases
B6
scratch
AA
Link 1


AA
Put out vectors including concepts from stems, clusters, and statistical phrases
A5
Put out sentences for syntactical analysis
A7
Put out data to print words not found
B1
B2
A
Yes
Is there more text to be looked up?
No
Link 1
(continued)
B1
B2
Print words not found
A3
Link 2
B5
grammar
A7
sentences
Perform syntactic analysis
B1
scratch
B2
sentence trees
Links 3-6
B2
sentence trees
B5
criterion trees
Perform graph matching to find syntactic phrases
A7
phrase vectors
Link 7
A5
stems, clusters, and statistical phrases
A7
syntactic phrases
A4
titles and citations
Merge and weight requests looked up during this run, if any
A3
print of vectors
B4
punched vectors
Link 8
Are correlations to be performed now?
Yes
No
Stack requests in core
Put out vectors
B1
BB


BB
B
A2
A6
previously looked up requests
Weight requests previously looked up or expanded by concept-concept or hierarchy
A3
print of vectors
A4 or B6
titles and citations
B1, B2, B4, B6, A5, A7, or A8
expanded requests*
Correlation now?
Yes
No
Put out vectors
B1
Stack requests in core
C
A5
stems, clusters, and statistical phrases
Merge and weight texts looked up during this run, if any
A3
print of vectors
A7
syntactic phrases
B4
punched bectors
A4
titles and citations
Correlations now?
Yes
No
Correlate text with requests
B2 or B6
correlations
Put out vectors
B1
Put out texts for document-document correlation
B1 or B6
C
Yes
More texts looked up during this run?
No
Expansion of requests only?
Yes
E
No
CC

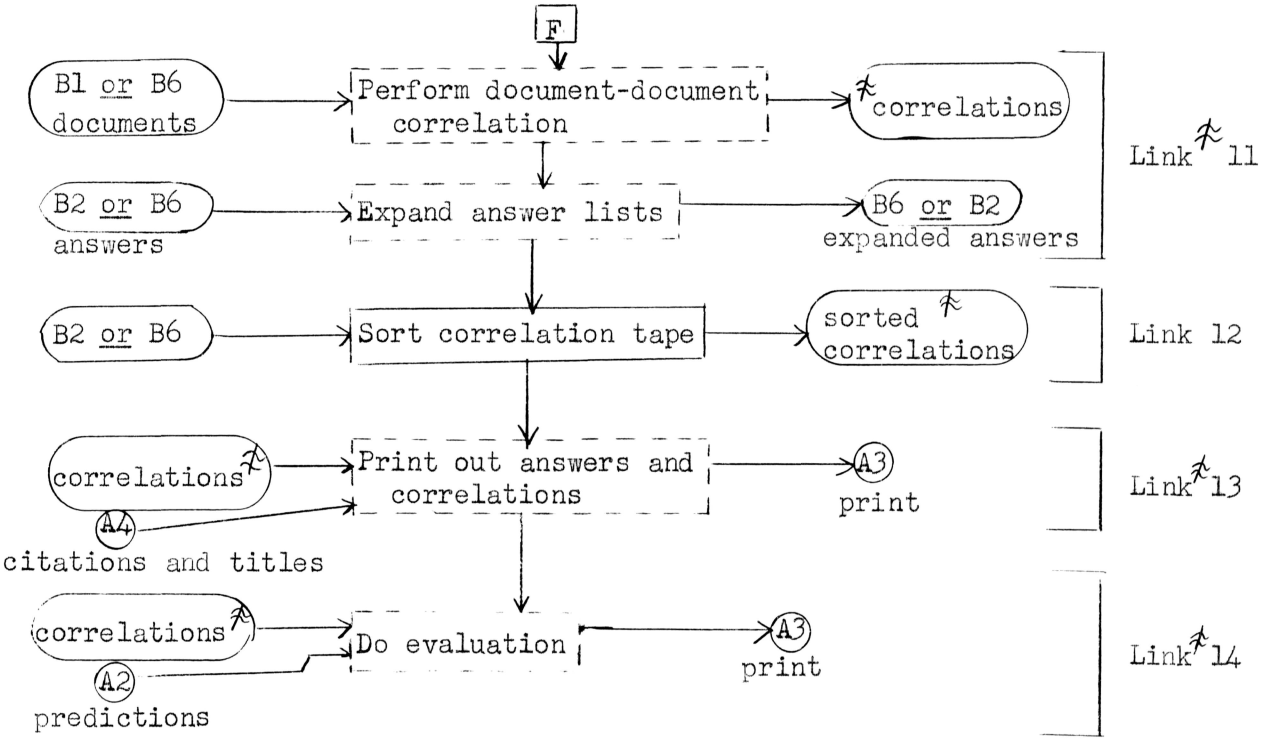
Link 8 (continued)

*The input tape for expanded vectors could be any of the above, depending on the output tape of the sort used for matrix inversion.


CC
D
A2, A6
previously looked up texts
B1, B2, B3 A9, A5, A7, or A8
expanded texts
Weight texts previously looked up or expanded this run
A3
print of vectors
B4
punched vectors
Correlation now?
Yes
No
Correlate text with requests
B2 or B6
correlations
Put out vectors
B1
Put out vectors for document-document correlation
B1 or B6
D
Yes
More texts?
No
No
Concept-concept or hierarchy expansion?
F
Yes
E
B1
vectors
Invert matrix and do concept-concept correlation
‡ correlations
Link 9‡
correlations‡
B5
hierarchy
Perform hierarchy and concept-concept expansion
‡expanded vectors
Link 10‡
D
Link 8 (continued)

† See note, preceding page.

‡ Intermediate and output tapes for all following links cannot be specified because of high-speed tape sort.



END

*See note, preceding page.

4. TAPE Assignments

Running on its full and most efficient configuration, SMART uses eight tapes on A channel and six tapes on B channel. However, perusal of the charts below will show that some tapes can be left off under certain circumstances when it is necessary to do so.

A. FORTRAN Tapes, Always Needed

A1 – FORTRAN monitor

A2 – input tape

A3 – output print tape

B3 – chain tape

B4 – output punch tape

B. SMART System Tapes

A6 – collection tape, only needed if permanent collection is used

B5 – library tape, needed if lookup or HIER options are performed

C. SMART Internal Tapes and Use

A4 – needed for lookup, for ANSWER MEDIUM or LONG, for concept-concept expansion on documents only

A5 – needed for lookup, for CONCON, HIER, ANSWER, SCORES

A7,A8 – needed for SYNTAX, CONCON, HIER, ANSWER, SCORES, DOCDOC

B1,B2 – needed for words not found, SYNTAX, CONCON, HIER, ANSWER, SCORES, DOCDOC

B6 – needed for lookup, CONCON, HIER, DOCDOC, ANSWER, SCORES, PUNCH

D. Tape Usage Chart

Note: After link 8, the exact use of scratch tapes cannot be predicted, since the high speed sort leaves its output on different tapes depending on where the last merge pass ended.

Link #/Tape	FORTRAN Tapes A1,A2,A3 B3,B4	B5 Library Tape	A6 Permanent Collection Tape	B1	B2	B6	A4	A5	A7	A8
1. lookup, stat. phrases, clusters		X		not found words X	X	scratch X	titles and citations X	vectors X	sentences X	
2. words not found				X	X		|	|	X	X
3. BTOKUN		X		X			|	|		X
4. SETUP		X		X	X		|	|		X
5. SYNTAX		X		X	X		|	|		X
6. EDIT		X		X	X		|	|		X
7. CRITER		X		scratch X	X		|	|	vectors X	X
8. merge, weight, punch and correlate vectors	always needed		X	vectors X vectors X	col-lections X	vectors or col-lections X	X X	X	X	
9. CONCON				| X	X	X	|	X	X	X
10. HIER and expand from CONCON		X		| X	X	X |	|	X	X	X
11. DOCDOC				X	X	X	|	X	X	X
12. SORT				X	X	X	|	X	X	X
13. REPORT				X	X	X	X	X	X	X
14. SCORES				X	X	X		X	X	X

X = used in this link | = in use during link

E. Special Tape Usage Chart

The chart which follows shows the use of tapes at the end of links 8, 9, and 10. The actual channel and unit of tapes X, Y, and Z cannot be

Type of Expansion	Requests			Documents			All		
Link	8	9	10	8	9	10	8	9	10

CONCON Only	B1
	(req,docs)
		X
		B1
		(req,docs)
			Z- (reqs)
			B1-(docs)
				B1(req,docs)
				A4(req)
					Y(docs)
					X
						Z -(docs)
						A4(req)
							B1
							(req,docs)
								X
								Y(req,docs)
									Z
									(req,docs)

Both	↓	↓	↓	↓	↓		↓	↓	↓

HIER Only	B1
	(req,doc)
		—	Z - (req)
			B1-(doc)
				B1(req,docs)	—	Z-(docs)
						B1(req)
							B1
							(req,docs)
								—	Z
									(req,docs)

X = CONCON tape

Y = sorted tape

Z = merged tape

predicted in advance since the high speed tape sort chooses its own output tape. At the end of link 10, control passes back to link 8 for merging and correlating.