LEXICO Guide No. 6

CONCORDING

Richard L. Venezky
Nathan Relles
Lynne A. Price

Computer Sciences Technical Report #287

December, 1976

LEXICO Guide No. 6

CONCORDING

Richard L. Venezky
Nathan Relles
Lynne A. Price

Computer Sciences Department
University of Wisconsin
December, 1976

TABLE OF CONTENTS

1. Introduction

This guide describes the options that a user may exercise in generating concordances and the manner in which he makes these specifications. Many of the commands may be entered in a CREATE or UPDATE block. In this case, they become collection defaults and need not be specified within each CONCORD block. For further discussion of defaults, see Guide 3, Section 2.

As a concordance is generated, each word in the text is tested to determine how it should be entered. In general, a word designated as a stopword is entered into the concordance only with its frequency. Words designated as keywords are entered along with each citation in which they occur. A user may provide the following information to be used in distinguishing keywords and stopwords:

(1) a stopword list--a list of words to be treated as stopwords;

(2) stop characters--characters which, when they appear at the beginning of a word, designate the word as a stopword;

(3) stoppable lengths--designation of words of a particular length as stopwords; and

(4) reversability--specification that all words that would otherwise be treated as stopwords should be treated as keywords and vice versa.

The user may also declare

(1) removable characters--characters to be stripped off the beginning, middle, or end of a word before it is entered into a concordance;

(2) a collating sequence for ordering keywords and stopwords in the concordance; and

(3) physical devices other than the printer to which concordance output should be directed.

The order of the operations performed on a word as it is entered into a concordance is described below in Section 5.

2. The CONCORD Block

LEXICO concords texts off-line. A concordance of a text that has previously been added to a collection, and perhaps been edited, is initiated by a block of the following form:

```
CONCORD text ;
   command 1 ;
   command 2 ;
       :
END ;
```

Here text is the name or code of a text (in the current collection) and the commands are concordance specifications that differ from the collection defaults. Concordance specifications are described in Section 3. An alternate block header is

```
CONCORD ;
```

If this shorter form is used, LEXICO asks WHICH TEXT? Either the text name or text code may be given in response.

One or more texts may be concorded as they are added to the collection, using the following block:

```
ADDCONCORD  text ;
   command 1 ;
   command 2 ;
       :
END ;
```

Where text is a text name and the commands are text input or concordance specifications entered in any order. In an ADDCONCORD block, declarations may be entered to describe the format of text input. See Guide 4 for detailed descriptions of these commands. This block also has an alternate block header,

```
ADDCONCORD ;
```

When it is entered, LEXICO prompts TEXT NAME, PLEASE? As described in Guide 4, the name entered in response must be the same as the name on the first line of the text input. If more than one text is to be added and concorded in an ADDCONCORD block, text must be the name of the first text in the group.

3. Concordance Specifications

3.1 Removable Characters

The user may specify one or more characters to be stripped from a word before the word is entered into a concordance. These characters may be declared as collection defaults in a CREATE or UPDATE block, or as text defaults in an ADD, CONCORD, or UPDATE TEXT block. The commands to remove characters from, respectively, the beginning of a word, end of a word, or anywhere within a word are

```
FRONTSTRIP characters ;
ENDSTRIP characters ;
SQUEEZE OUT characters ;
```

where characters is a string of up to 64 removable characters, entered in any order without intervening blanks. If the string spells a reserved word or contains a reserved character (See Guide 2, Section 3.4) the entire string must be enclosed in single quotation marks.

These commands do not affect the text as it is stored in the collection, or the spellings of the words in the citations listed with each keyword. Nor are they cumulative; if two SQUEEZE OUT commands, for example, specifying two different sets of characters, are entered, only the characters given in the second command will be removed from words in a text as the words are entered into a concordance.

If the first character of a word is a frontstrip character, it is removed and the testing process is repeated on the next character. This continues until the first non-frontstrip character is found.

This same procedure applies, mutatis mutandis, to endstrip.

As an example of the use of these commands, suppose capital letters are prefixed by the symbol '>'. Thus, when text input is prepared, 'The' will be represented by >THE. The command

```
FRONTSTRIP > ;
```

can be used to ensure that each word type in a text will be entered into a concordance only once, even if it appears in both capitalized and uncapitalized forms. In this case, if "the" is a keyword, all citations containing both >THE and THE will be grouped together under THE. Without the frontstrip declaration, there would be an entry for >THE and another for THE.

When an earlier declaration of removable characters is to be cancelled, the commands

```
DONOT FRONTSTRIP ;
DONOT SQUEEZE OUT ;
DONOT ENDSTRIP ;
```

may be used.

The removable characters in effect in any block are included in the display produced by the command

```
SPECS ;
```

No removable characters are included in the system defaults.

3.2 Stopwords

Unless the user defines stopwords, whenever a text is concorded, all words will be keywords. Declarations pertaining to stopwords may be entered as collection defaults in a CREATE or UPDATE block, or as temporary text values in a CONCORD or ADDCONCORD block.

The user may associate with each collection a list of stopwords to be used when concordances are generated. The user may create his own list of words or use LEXICO's standard stopword list. In either case, he may subsequently add or delete words from the list. The maximum number of words the list may contain varies with their length; approximately 200 words averaging six letters each may be entered. The collection's stopword list is defined in a CREATE or UPDATE block. It may be temporarily modified for an individual text in a CONCORD or ADDCONCORD block.

The command

STANDARD STOPWORDS ;

is used to select the system-supplied stopword list (which is given in Section 7). The command

ADD STOPWORDS word1 word2 word 3... ;

adds the specified words to the existing stopword list. As discussed in Guide 2, Section 3.4, words containing reserved characters (e.g., blank, comma, semicolon) and reserved words that have special meaning to the LEXICO system (e.g., CREATE, INPUT, ON, OF) must be enclosed in single quotes. Any stopword may be enclosed in quotes.

The similar command

DELETE STOPWORDS word1 word2 word3... ;

is used to remove the specified words from the existing stopword list.

If an error (e.g., an attempt to delete a word that is not on the stopword list) is made in an ADD STOPWORD or DELETE STOPWORD command, LEXICO will process the correct portions of the command.

The command

DELETE ALL STOPWORDS ;

erases the existing stopwords list. Finally, the stopword list may be displayed by the command

SHOW STOPWORDS ;

The user may also declare stop characters and stoppable lengths. The command

AUTOSTOP characters ;

(where characters is a string of up to 64 characters, enclosed in quotes if necessary --see Guide 2, Section 3.4-- and listed without intervening blanks, in any order) is used to indicate that all words beginning with one of the letters in the string characters should be treated as stopwords. A declaration of stop characters can be nullified with the command

DONOT AUTOSTOP ;

The commands

STOP LENGTH n ;

and

STOP LENGTH n OR LESS ;

are used to indicate, respectively, that all words of length n and that all words no longer than n characters should be treated as stopwords. Only one stoppable length may be in effect at any one time.

The command

REVERSE ;

signifies that the rules defining keywords and stopwords should be interchanged. If this command is entered an even number of times, the stopword list, stop characters, and stoppable lengths, if any, are used to recognize stopwords; if the command is entered an odd number of times, these concordance specifications refer to keywords.

Stop characters, stoppable lengths, and an indication of whether the stopword rules actually refer to stopwords or to keywords are all included in the display produced by the command

SPECS ;

3.3 Collating Sequence

Concordance entries are alphabetized according to a collection's collating sequence. Unless a different sequence is specified in a CREATE or UPDATE block, the standard collating sequence is used. This is

@[]#^ ␣ABCDEFGHIJKLMNOPQRSTUVWXYZ)-+<=>&$*(%:?!,\0123456789';/."

where ␣ denotes the blank symbol. A different collating sequence may be temporarily defined for an individual text in a CONCORD or ADDCONCORD block.

The command

COLLATE NEW 'sequence' ;

is used to define a new collating sequence. The user need specify only the characters that appear in his texts; LEXICO will insert omitted characters at the end of the sequence. Users who define their own collating sequence will almost always include the blank symbol as the first character. This is done so that whenever a word has a prefix which is also a word (e.g., as the prefixes there), the shorter word (in this case, the) will precede the longer one (there) in the concordance.

LEXICO also provides an Old English collating sequence:

ƀ/(A$BCDE-FGHIJKLMNOPQRST*+UVWXYZ0123456789@[]#^)<=>&%:?!,\';."

where ƀ, $, -, *, + are used to denote respectively the blank symbol, æ, ę, þ, and ð.

This is selected with the command

COLLATE OE ;

The standard collating sequence can be restored with the command

COLLATE STANDARD ;

The collating sequence in effect in any block is included in the display produced by the command

SPECS ;

3.4 Delimiters

Delimiters are entered as collection defaults in a CREATE or UPDATE block and are displayed by the SPECS command. A complete discussion of all LEXICO delimiters appears in Guide 4, Section 4. However, two types of delimiters are especially pertinent to concording. It may be necessary to change word delimiters for a text with unusual punctuation. Up to six word delimiters may be specified as text defaults in an ADD, CONCORD, ADDCONCORD, or UPDATE TEXT block with the command

WORD DELIMITERS d1 d2 ... ;

One or two note delimiters may be specified as temporary text values in an ADD, CONCORD, or ADDCONCORD block using the command

NOTE DELIMITERS d1 d2 ;

In some cases, such as when notes consist of translations, it may be desirable to concord both a text and its notes. If the original text input were properly prepared, this could be done by concording the text once, changing the note delimiters, and re-concording.

3.5 Concordance Output

Whenever a text is concorded, LEXICO generates a concordance in two forms: a printable version, which usually is printed as the output of the concordance run, and an internal one, usually stored in the collection. The latter may be used to generate slips after headword classification has been completed.

Users may have the printable version written to a file (and not printed) with the command

CONCORDANCE OUTPUT ON file ;

where file is a file name of up to twelve characters, including any qualifier, optionally including a terminal period, and perhaps enclosed in quotes. This specifies that the concordance should be directed to a file whose name is file. This file may then be printed, transcribed onto microfiche or written onto magnetic tape. It must be created outside of LEXICO (see Guide 2, Section 8.2). This command affects only the printable version of the concordance; the internal version will be stored in the collection.

Both versions of the concordance may be directed to magnetic tape with the command

CONCORDANCE OUTPUT ON TAPE ;

This will cause two SDF files to be written at high density (800 bpi) in odd parity in @COPY,G format. The first file is the printable version; the second file is the internal one. LEXICO will prompt for the MACC tape number of the tape to be used and the number of files that already exist on that tape. The user must be very careful to supply this information correctly as the system

does no checking for possible errors. If the user enters this data erroneously, a repeated CONCORDANCE OUTPUT ON TAPE command will cause LEXICO to prompt for this information again. Of course, if an error is noticed after the END statement has been entered, the user may indicate the run should be cancelled when the system asks for a priority (see Guide 2, Section 7). When slips are generated, LEXICO will need to read the internal concordance back from tape (see Guide 7, Section 6.2). It should be mentioned that the daily file charge for a collection can be greatly reduced if this internal concordance is stored on tape. For a brief discussion of the use of tapes with LEXICO, see Guide 2, Section 8.

The system default for the concordance output device may be restored with the command

CONCORDANCE OUTPUT ON PRINTER ;

This option is included in a display produced by the SPECS command. It may be specified as a collection default in a CREATE or UPDATE block or as a temporary text value in a CONCORD or ADDCONCORD block.

3.6 Arrangements of Keywords and Stopwords

As part of the output from a CONCORD or ADDCONCORD block, the user may elect a list of the keywords and stopwords of a list, ordered by the frequency with which the words appear in the text. This is done with the command

LIST TYPES BY FREQUENCY ;

The words may also be listed in an order computed by reversing each word (spelling it backwards) and alphabetizing the result. This groups words with similar endings. The words are printed in normal form, however. This list is selected with the command

LIST TYPES REVERSED ;

The user who does not want these lists may use the command

DONOT LIST BY FREQUENCY ;

and

DONOT LIST REVERSED ;

These commands may be entered as collection defaults in a CREATE or UPDATE block, and as text defaults in an ADD, CONCORD, ADDCONCORD, or UPDATE TEXT block. The system default is not to generate the lists.

These lists may be obtained in separate off-line jobs, (see Guide 3, Section 4.6).

3.7 Displaying Concordance Specifications

Text input and concordance specifications (except the stopword list) are displayed with the command

SPECS ;

This may be entered in CREATE, UPDATE, ADD, CONCORD, ADDCONCORD, UPDATE TEXT, or EDIT blocks. It results in a display of all specifications pertinent to the block.

3.8 Text Memo

In any text-specific block, the statement

MEMO 'note about text';

may be entered, where note about text is a string of up to 100 characters. This memo will be displayed whenever the status of the text is displayed (see STATUS OF command in Guide 3, Section 4.2). One use of this statement is to enter text titles longer than the 12 characters allowed in text names.

4. Sample Concordance

Figure 1 shows the concordance that resulted from the following interaction:

```
>concord 1 ;
 TEXT 'JFK          '  (TEXT CODE:     1) AVAILABLE:
 YOU MAY ENTER CONCORDANCE SPECIFICATIONS:

>    standard stopwords ;
>    word delimiters blank quote ;
>    frontstrip > ;
>end ;
 CREATE BACKUP IMMEDIATELY BEFORE THIS PROCESS?  (Y OR N)
           _>n ;
 WHEN? (I, T, O, W)
           _>t

 RUN IDENTIFICATION: X01673        (SAVERN0910*6105649)
```

The following text has two-level IDs and was ADDed with semicolon and period as citation delimiters and square brackets as note delimiters:

```
(JFK/1)'>ASK NOT WHAT YOU CAN DO FOR YOUR COUNTRY ;
(JFK/2)ASK WHAT YOUR COUNTRY CAN DO FOR YOU'[>J. >F. >KENNEDY].
```

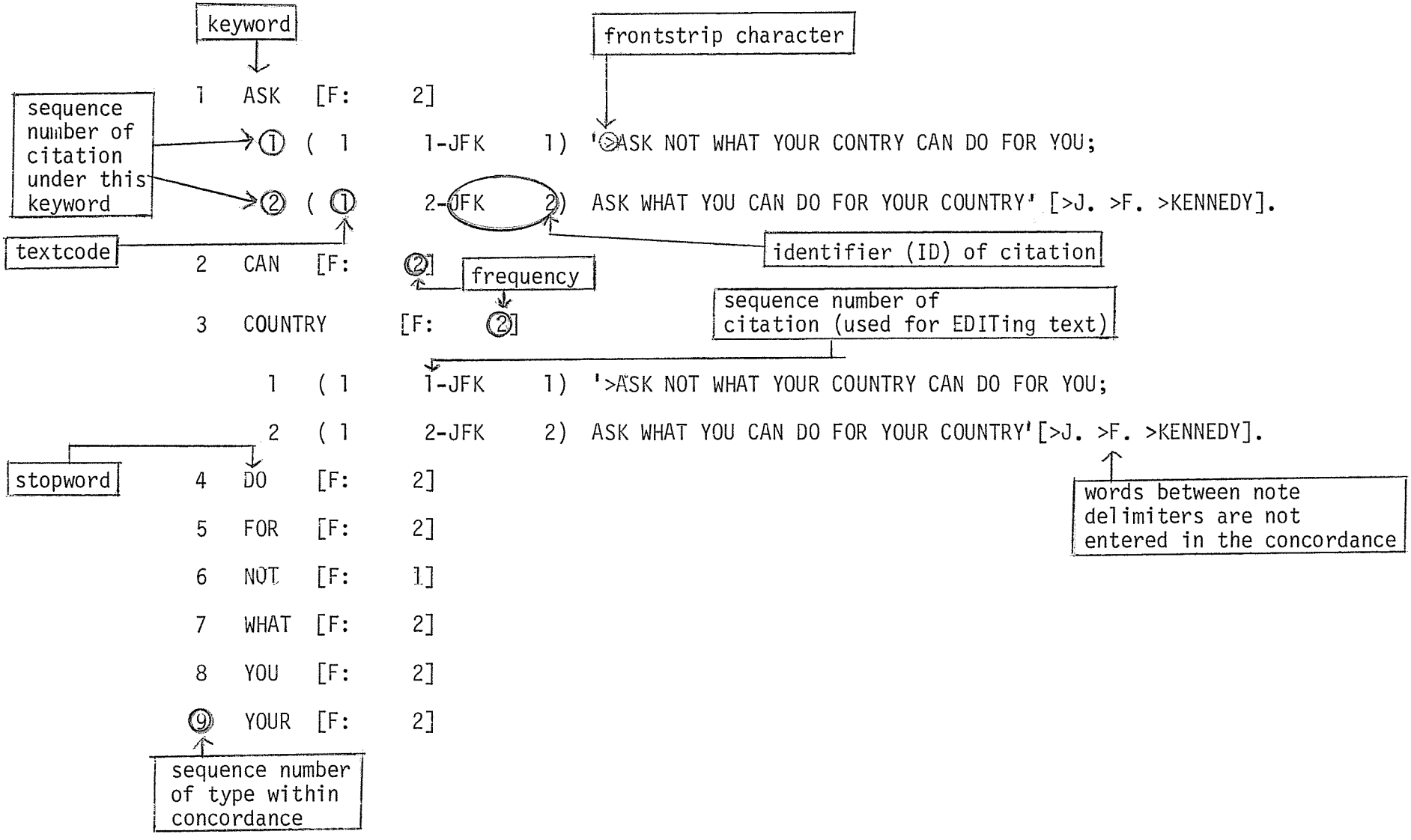

Figure 1

Sample Concordance

5. Order of Operations

The user may enter concordance specifications in any order. However, when actually generating a concordance, LEXICO performs the resulting tests and operations in a fixed order. In some cases, this order may affect the result. As a word is processed, removable characters are discarded before it is determined whether the word is a keyword or a stopword.

Squeeze out characters are removed first. Then all frontstrip characters are removed from the beginning and endstrip characters from the end of the word. If a word disappears entirely during these processes, a warning is printed and nothing is entered into the concordance.

6. Synonyms and Abbreviations

Members of the following groups of expressions may be used interchangeably when entering concordance specifications:

BEFORE, BY, EXCEPT, FOR, FROM, INTO, OF , ON, OR, OUT, THRU, TO, WITH ;
CONCORDANCE, CONCORDANCES ;
DELETE, D ;
DELIMITER, DELIMITERS, DELIM ;
DONOT, NO, NOT ;
NOTE, NOTES ;
SHOW, S, SH, DISPLAY ;
STOPWORD, STOPWORDS ;
TYPE, TYPES ;
WORD, WORDS ;

7. Standard Stopwords

The words in LEXICO's standard stopword list are:

A	FROM	MY	THERE
ABOUT	HAD	NEW	THESE
AFTER	HAS	NO	THEY
ALL	HAVE	NOT	THIS
ALSO	HE	NOW	THROUGH
AN	HER	OF	TIME
ANY	HIS	ONE	TWO
ARE	I	ONLY	UP
AS	IF	OR	WAS
AT	IN	OTHER	WAY
BACK	INTO	OUR	WE
BE	IS	OUT	WELL
BEEN	IT	OVER	WERE
BEFORE	ITS	SAID	WHAT
BUT	LIKE	SHE	WHEN
BY	MADE	SO	WHERE
CAN	MAN	SOME	WHICH
COULD	MANY	SUCH	WHO
DID	MAY	THAN	WILL
DO	ME	THAT	WITH
DOWN	MORE	THE	WOULD
EVEN	MUST	THEIR	YEARS
FIRST	MUCH	THEM	YOU
FOR	MOST	THEN	YOUR

8. Command Summary

8.1 CONCORD Block

The following commands may be entered in a CONCORD block:

```
SPECS ;

WORD DELIMITERS d1 d2 ...d6 ;
NOTE  DELIMITERS d1 d2 ;

CONCORDANCE OUTPUT ON PRINTER ;
CONCORDANCE OUTPUT ON TAPE ;
CONCORDANCE OUTPUT ON file ;

SQUEEZE OUT characters ;
FRONTSTRIP characters ;
ENDSTRIP characters ;
DONOT SQUEEZE OUT ;
DONOT FRONTSTRIP ;
DONOT ENDSTRIP ;

DELETE ALL STOPWORDS ;
ADD STOPWORDS word1 word2... ;
DELETE STOPWORDS word1 word2... ;
STANDARD STOPWORDS ;
SHOW STOPWORDS ;
STOP LENGTH n ;
STOP LENGTH n OR LESS :
AUTOSTOP characters ;
DONOT AUTOSTOP ;
REVERSE ;

COLLATE OE ;
COLLATE STANDARD ;
COLLATE NEW sequence ;

LIST TYPES BY FREQUENCY ;
LIST TYPES REVERSED ;
DONOT LIST TYPES BY FREQUENCY ;
DONOT LIST TYPES REVERSED ;

MEMO 'note about the text' ;

END ;
IGNORE ;
```

8.2 ADDCONCORD Block

The following commands may be entered in an ADDCONCORD block:

```
SPECS ;

WORD DELIMITERS d1 d2 ... d6 ;
ID DELIMITERS d1 d2 ;
ID LEVEL DELIMITER d1 ;
CITATION DELIMITERS d1 d2 ... d6 ;
NOTE DELIMITERS d1 d2 ;
TEXT DELIMITER d1 ;

SEQUENCE ON CITATION ;
SEQUENCE ON character1 ;
SEQUENCE ON character1 EXCEPT BEFORE character2 ;
SEQUENCE ON n CITATIONS ;
DO NOT SEQUENCE ;

INPUT ON CARD FILE file ;
PHYSICAL LENGTH n ;
LIMIT CITATIONS TO n ;

LIST TEXT ;
DONOT LIST TEXT ;
TEXT OUTPUT ON PRINTER ;
TEXT OUTPUT ON TAPE ;
TEXT OUTPUT ON file ;

CONCORDANCE OUTPUT ON PRINTER ;
CONCORDANCE ON TAPE ;
CONCORDANCE OUTPUT ON file ;

SQUEEZE OUT characters ;
FRONTSTRIP characters ;
ENDSTRIP characters ;
DONOT SQUEEZE OUT ;
DONOT FRONTSTRIP ;
DONOT ENDSTRIP ;

DELETE ALL STOPWORDS ;
ADD STOPWORDS word1 word2 ... ;
DELETE STOPWORDS word1 word2 ... ;
STANDARD STOPWORDS ;
SHOW STOPWORDS ;
STOP LENGTH n ;
STOP LENGTH n OR LESS ;
AUTOSTOP characters ;
DONOT AUTOSTOP ;
REVERSE ;

COLLATE OE ;
COLLATE STANDARD ;
COLLATE NEW sequence ;

LIST TYPES BY FREQUENCY ;
LIST TYPES REVERSED ;
DONOT LIST TYPES BY FREQUENCY ;
DONOT LIST TYPES REVERSED ;

MEMO 'note about the text' ;
END ;
IGNORE ;
```

9. Reserved Words

9.1 The CONCORD Block

The following words are reserved in a CONCORD block:

ADD
ADDCONCORD
ALL
AUTOSTOP
BACKUP
BASETYPE
BEFORE
BLANK
BY
CARD
CARDS
CHARACTER
CHARACTERS
CIT
CITATION
CITATIONS
CLEAR
CLEANUP
COLLATE
COLLECTION
COLON
COMMA
CONCORD
CONCORDANCE
CONCORDANCES
CREATE
D
DELETE
DELIM
DELIMITER
DELIMITERS
DIR
DIRECTORY
DISPLAY
DO
DONOT
EDIT
END
ENDSTRIP
EXCEPT
FILE
FOR
FREQUENCY
FROM
FRONTSTRIP
ID
IGNORE
INPUT
INTO
LENGTH
LESS
LEVEL
LIMIT
LIST
LISTS
LOOKUP
MEMO
MINUS
NEW
NO
NOT
NOTE
NOTES
OE
OF
ON
OPTION
OR
OUT
OUTPUT
PACK
PHYSICAL
RENAME
RESPELL
RESPELLED
RESTORE
REVERSE
REVERSED
ROUTE
RULE
RULES
S
SEMI
SEQUENCE
SH
SHOW
SLIPS
SPECS
SPELLING
SQUEEZE
STATUS
STANDARD
STOP
STOPWORD
STOPWORDS
TEXT
TEXTS
THRU
TO
TYPE
TYPES
UPDATE
WITH
WORD
WORDS

9.2 The ADDCONCORD Block

The following words are reserved in an ADDCONCORD block:

ADD
ADDCONCORD
ALL
AUTOSTOP
BACKUP
BASETYPE
BEFORE
BLANK
BY
CARD
CARDS
CHARACTER
CHARACTERS
CIT
CITATION
CITATIONS
CLEAR
CLEANUP
COLLATE
COLLECTION
COLON
COMMA
CONCORD
CONCORDANCE
CONCORDANCES
CREATE
D
DELETE
DELIM
DELIMITER
DELIMITERS
DIR
DIRECTORY
DISPLAY
DO
DONOT
EDIT
END
ENDSTRIP
EXCEPT
FILE
FOR
FREQUENCY
FROM
FRONTSTRIP
ID
IGNORE
INPUT
INTO
LENGTH
LESS
LEVEL
LIMIT
LIST
LISTS
LOOKUP
MEMO
MINUS
NEW
NO
NOT
NOTE
NOTES
OE
OF
ON
OPTION
OR
OUT
OUTPUT
PACK
PHYSICAL
RENAME
RESPELL
RESPELLED
RESTORE
REVERSE
REVERSED
ROUTE
RULE
RULES
S
SEMI
SEQUENCE
SH
SHOW
SLIPS
SPECS
SPELLING
SQUEEZE
STATUS
STANDARD
STOP
STOPWORD
STOPWORDS
TEXT
TEXTS
THRU
TO
TYPE
TYPES
UPDATE
WITH
WORD
WORDS