DOCUMENT RESUME

ED 108 633 IR 002 150

AUTHOR Porch, Ann
TITLE Design Document: KWIC Module; L.A.P. Version I.
INSTITUTION Southwest Regional Laboratory for Educational Research and Development, Los Alamitos, Calif.
REPORT NO SWRL-TN-5-72-37
PUB DATE 26 May 72
NOTE 9p.

EDRS PRICE MF-$0.76 HC-$1.58 PLUS POSTAGE
DESCRIPTORS *Computer Programs; Content Analysis; Design; Electronic Data Processing; *Indexes (Locaters); Item Analysis; *Language Research; Language Usage; *Morphology (Languages); Permuted Indexes; *Specifications; Systems Development
IDENTIFIERS Computer Software Specifications; *Language Analysis Package; LAP

ABSTRACT

The Language Analysis Package (LAP) was developed by the Southwest Regional Laboratory (SWRL) to assist researchers in the analysis of language usage. The function of the KWIC (Keyword-in Context or Concordance) Module of the LAP is to produce keyword listings from the input text being analyzed. Such listings will contain location information broken down by document identifier, page, paragraph, and line. Other design features are presented in this document together with the file layout specifications of the program's output. (DGC)

SOUTHWEST REGIONAL LABORATORY TECHNICAL NOTE

DATE: May 26, 1972

NO: TN 5-72-37

TITLE: DESIGN DOCUMENT: KWIC MODULE - L.A.P. VERSION I

AUTHOR: Ann Porch

ABSTRACT

This is one of a series of technical design specificiations for individual modules of the Language Analysis Package (L.A.P.)

The KWIC Module will provide the user with a KWIC index including context and location information. The user will be able to specify length of context desired and either of two basic formats for output.

DESIGN DOCUMENT: KWIC MODULE - L.A.P. VERSION I

The following document will serve as design specifications for programming. It is one of a series of technical design specifications for individual modules of the Language Analysis Package (L.A.P.)[1]. The section of the system design to which it is related is 6.2.0.

Program Objective

The function of the KWIC Module will be to produce Keyword-in-Context (Concordance) listings from the input text. Such listings will contain location information which will be broken down into four main categories: level 1 (ex. Document ID), level 2 (ex. Page), level 3 (ex. paragraph), and level 4 (ex. line). The KWIC Module will usually be run in conjunction with the List Search Module and inclusion or exclusion dictionaries. It may also run in conjunction with the Sensitive Module with user defined limitations on the portions of the input data to be processed.

Constraints and Limitations

There will be no constraints or limitations for input data for the KWIC Module, other than those described in design specifications for the Scan Module.[2]

---

1. See Porch, Ann, TM 5-72-06, "Language Analysis Package (L.A.P.) Version I System Design."
2. See Porch, Ann, TN 5-72-27, "Design Document: Scan Module - L.A.P. Version I."

Options and Defaults

Options and defaults for the KWIC Module will be as follows:

- Length of before context. The length of context will be specified in number of characters. (Default = 48)

- Length of keyword (Default = 16)

- Length of after context. The length of context will be specified in number of characters. (Default = 48)

- Form for identifying location information within the input text. (Default=@TXT for level 1; @PG for level 2; @SEC for level 3; and a computer generated count of input records (lines) for level 4)

- Sort field and ordering (Default = keyword, location in ascending order)

- Output format. There will be two choices: A.) one line, keyword centered, and, B.) two line, keyword right justified on line 1. (Default = one line, keyword centered)

Data File Specifications

Input

Data will be input to the Scan Module and will be considered as a stream of characters. No notice will be taken of record boundaries. Data which has been prepared utilizing other conventions can be handled by use of the TRANSLATE Module. The SCAN Module will be used in conjunction with the KWIC Module. It will break the text into words and check for special characters needed by sub-routines handling location information, etc. Parameters passed from SCAN to

KWIC will be text array identification, and the beginning and end point of the word.

Output

Output data will fall into one of two user specified formats. An example of each is shown in Appendix A. Output may be obtained on the high speed printer, punched cards, or magnetic tape.

Significant Algorithms

There are three significant algorithms connected with the KWIC Module. They are:

- The Three Record, Circular Read Algorithm (described more fully in "Design Document: Scan Module - L.A.P. Version I")
- The Contexting Algorithm
- The Location Information Algorithm

Three Record, Circular Read Algorithm. To find the context, the specified length of the before context will be subtracted from the pointer value for the beginning of the word. The result will be stored in a variable indicating the beginning point for the context.

To find the remaining portion of the context, the after context length will be added to the pointer value for the end of the word. The result will be stored in a variable indicating the ending point for the context.

A test will be made to determine if special action needs to be taken because values of either of the variables indicating beginning point and ending point for the context falls outside the bounds of the array. If so, special action will be taken as described later

below. If both variable values fall within the bounds of the array, the context may be output directly from the array itself to a file for later sorting. A standard, implied-Do type print statement may be used.

If the value of the variable indicating the beginning point of the context is negative, special handling is required. The negative value will be added to the upper bound of the array and the result stored in the begin-context variable. When the context is output to the file for later sorting, the print statement will have two parts. First the array values from the value of begin-context variable to the end of the array will be output. Then the array values from position 1 of the array to the value of the end-context variable will be printed.

If the value of the variable indicating the ending point of the context is greater than the upper bound of the array, special handling is also required. The array's upper bound value will be subtracted from the value of the end-context variable and the result stored in the end context variable. Again, when the context is output to the file for later sorting, the print statement will have two parts. As before, first the array values from the value of the begin-context variable to the end of the array will be output. Then the array values from position 1 of the array to the value of the end-context variable will be printed.

The Location Information Algorithm. Variables will be set up for each of the following four levels of location information:

- Level 1 (ex. document ID)
- Level 2 (ex. page)
- Level 3 (ex. paragraph)
- Level 4 (ex. line)

The user will specify a flag character (which must not be the same as any of the characters he has specified as valid word builders) and up to four following characters to indicate the function he is indicating. For example, the user may specify that all location information will be flagged with an '@'. He may flag pages with "@PG" and Paragraphs "@PAR".

The SCAN Module will pass to the Location subroutine the information that it has found a special character. The Location subroutine will then check to see if the special character found matched those which the user has specified as meaningful for the KWIC Module. If so, special action will be taken. Variables will be set up for values for each level. Level 1 (ex. document ID) identifiers will consist of three alphanumeric characters supplied from the input stream. (Default - TXT) All other variables will be counters. Counts for Level 2 indicators will be dependent on Level 1, and will be restarted whenever a new level 1 indicator is found. Likewise, counts for level 3 and 4 indicators will be dependent on level 2, and be restarted whenever a new level 2 is encountered.

For output, the contents of the four level variables will be introduced into the output stream in the positions appropriate to the

output format option chosen by the user.

Significant Variables

There are four significant variables in the KWIC Module. They are:

-Three Record Array

-Pointers to word beginning and end

-Pointers to context beginning and end

-Four variables for the levels of location information

Error and other Messages

The following messages are printed out by the KWIC Module:

- "End of Job" if the run terminates normally when there are no further texts to be processed.

Called by and/or Calls

The KWIC Module is called only by the Control Module.

The KWIC Module may call the following other modules:

- List Search (5.1.0) optional
- Sensitive (5.2.0) optional
- Sort (4.5.0) always called

SWRI

COMPUTER SYSTEMS
FILE/RECORD LAYOUT

TITLE: _____ DATE: _____

FILE ID: _____

PROGRAMMER: _____

RECORDING MODE:
FIXED ☐
LENGTH _____
VARIABLE ☐
MAX _____
MIN _____

BLOCKING FACTOR:
BLOCK CONTAINS
____ RECORDS

REMARKS:

Appendix A
Output

CARD LAYOUT ☐

TAPE
RECORD LAYOUT ☐

FORMAT LEGEND:

△ = Blank	P = Packed Dec.
A = Alpha	O = Octal
N = Numeric	H = Hexadecimal
X = Alpha/Num.	B = Binary

RECORD I.D.
Output option A

location before context Keyword after context

FORMAT
POSITIONS 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100

☐ CONTINUED

RECORD I.D.
Output option B

location before context Keyword

FORMAT
POSITIONS 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100

☐ CONTINUED

RECORD I.D.
Output Option B cont.

after context

FORMAT
POSITIONS 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100

☐ CONTINUED

RECORD I.D.

FORMAT