A Computer Program to Index or Search Linear Notations*†

KURT D. OFER
Scientific Department, Ministry of Defence, Tel-Aviv, Israel

Received May 21, 1968

The program will select from a file those notations which contain specified Boolean expressions of notation symbols; a wide range of selection criteria is available and batching of questions is possible. The program can be used to prepare a Key Symbol Out of Context index, to disseminate notices on additions to the file to interested chemists, to search the file for compounds with certain structural elements and list or/and count the selected items, or to convert from a linear to a fragmentation code. Although the program was designed to process Wiswesser line notations, it can be adapted to other linear notations with trivial changes; it is coded in FORTRAN and can thus be implemented on almost any computer.

As a rule, programs for processing notations are written in assembler language; the reason for this is that most problem-oriented languages cannot conveniently handle the character strings which form a notation. It takes a lot of time to code and debug an assembler program; it can be run only on a specific machine and is not easily modified when requirements change, as they are bound to do in the experimental phase of the work. PL/I and certain dialects of FORTRAN IV (notably those implemented on CDC computers) have character-handling capability together with all the other advantages of high-level languages, but they are handicapped by their dependence on a certain computer series and its character set.

The program to be described here is written in FORTRAN II and can therefore be run on almost any computer; symbols having a special meaning are defined by an input card, thus the implementation does not depend upon the character set used. Indeed, the same program is able to process several files of the same notation punched in different character sets; since it is easily changed, it should be ideal for experimental work such as:

1. Investigating the efficacy of classification numbers[2] or other sieves.
2. Obtaining statistics of the file.
3. Comparing modifications of the notation syntax (such as the effect of methyl contraction or multipliers).
4. Defining groups for estimation of physical properties.[3]

It can, of course, also be used for production runs until a more efficient special-purpose program is available. The program has been run satisfactorily on a Philco 2000 and on a CDC 3400 without any change (as a matter of fact, it was originally coded in ALTAC and then automatically translated into FORTRAN by a program written for this task).

There are many uses for a notation file or parts thereof (such as periodical additions), the main ones being:

a. To search the file for notations with specified terms (retrieval), the selected items to be listed and/or counted.
b. To index the file.
c. To disseminate notices on additions to the file.
d. To prepare a file in a different notation such as a fragmentation code (mainly in installations where an earlier fragmentation file is in use).

The common denominator of all these processes is the selection of records from the file according to specified selection criteria. The output takes different forms, printed listings or magnetic tapes (or disks) for subsequent sorting (the program, as written, provides for printed output only; the change to writing on tape or disk is trivial and most installations have their own sorting routines). The file is supposed to be on cards with the following format: C/c 1-6: Serial number, c/c 7-80: Notations. Up to two trailer cards (same serial number, no other indication required) are allowed, thus notations with a maximum length of 223 symbols can be processed. Change to other formats or to tape input is again trivial. Information separated from the notation by more than two delimiter symbols (normally blanks) will be treated as comment and not processed but will appear in the printed output.

The selection criteria have evolved from those described earlier[4] in connection with a Key Symbol Out of Context (KSOC) program. They are specified by cards at the start of each run, preceded by a definition card of the format given in Table I. The definitions of the selection

* Presented before the Division of Chemical Literature 155th Meeting, ACS, San Francisco, Calif., April 4, 1968.

† The Program deck is available from Dr. H. T. Bonnett, G. D. Searle & Co., P.O. Box 5110, Chicago, Ill. 60680

Table I. Lay-Out of Definition Card

C C	Definition	Example
1)	Numeral 1	1
2)	Numeral 2	2
3)	Numeral 3	3
4)	Numeral 4	4
5)	Numeral 5	5
6)	Numeral 6	5
7)	Numeral 7	7
8)	Numeral 8	8
9)	Numeral 9	9
10)	Numeral 10	0
11)	General Numeral (Alkyl)	A
12)	Delimiter	blank
13)	End of profile	≠
14)	Halogen 1	E
15)	Halogen 2	F
16)	Halogen 3	G
17)	Halogen 4	I
18)	Halogen 5	J
19)	Negation	Downward arrow
20)	Start of OR bracket	(
21)	End of OR bracket	)
22)	End of term in OR bracket	,
23)	"Don't care" symbol	=
24)	Switch symbol	Dollar
25)	Profile in profile	Asterisk
26)	Search for Alkyl	
27)	Search for Halogen	

Table II. Definitions of Selection Criteria

⟨Switch identifier⟩ ::= Numeral 1 | Numeral 2 | . . . . | Numeral 9

⟨Search identifier⟩ ::= any character which is not a ⟨Switch identifier⟩

⟨Search operator⟩ ::= Numeral 1

⟨Close-switch operator⟩ ::= Numeral 2 Will close the designated switch.

⟨Open-switch operator⟩ ::= Numeral 3 Will open the designated switch.

⟨Profile identifier⟩ ::= ⟨Search identifier⟩ ⟨Search operator⟩ | ⟨Switch identifier⟩ ⟨Close-switch operator⟩ | ⟨Switch identifier⟩ ⟨Open-switch op.⟩

⟨Operator symbol⟩ ::= Delimiter | End of profile | Negation | Start OR | End OR | End OR term | Switch symbol | Profile in profile | Search for Alkyl | Search for Halogen | Don't care

⟨Notation symbol⟩ ::= any character which is not a ⟨Operator symbol⟩

⟨Notation character⟩ ::= ⟨Notation symbol⟩ | Don't care

⟨Primitive term⟩ ::= ⟨Notation character⟩ | ⟨Notation character⟩ ⟨Primitive term⟩ | Profile in profile ⟨Search identifier⟩ | Switch symbol ⟨Switch identifier⟩ 'Profile in profile' will retrieve any predefined profile which is answered by the notation, 'Switch' will be a hit if the designated switch is closed.

⟨OR term⟩ ::= ⟨Primitive term⟩ End OR term | ⟨OR term⟩ ⟨OR term⟩

⟨OR bracket⟩ ::= Start OR ⟨OR term⟩ End OR Will retrieve any of the OR's

⟨Term⟩ ::= ⟨Primitive term⟩ | ⟨OR bracket⟩ | ⟨Term⟩ ⟨Term⟩ | ⟨Term⟩ Negation ⟨Term⟩

⟨Profile⟩ ::= ⟨Profile identifier⟩ ⟨Term⟩ End of profile

⟨Notation⟩ ::= ⟨Notation symbol⟩ Delimiter Delimiter Delimiter | ⟨Notation symbol⟩ ⟨Notation⟩

criteria in Backus Normal Form[1] are detailed in Table II. These cards are followed by a card with a delimiter symbol in c/c 1; after this come the file cards, the end-of-file being signalled by a card with zero (or negative) serial number. The program will print the selection criteria, then—for each selected notation—the profile identification (unless this is a switch designation), the multiplicity of this profile in the notation, the serial number, and the full notation including any comments (see above). At the end-of-file signal, the number of notations answering each profile is printed.

The program as written allows the simultaneous processing of up to 100 profiles with not more than 1000 characters; it occupies (together with the compiler-generated input-output routines) 6090 words of memory in the CDC 3400. The processing speed is limited by the input-output, which is not overlapped in this machine. The 10 notations and 15 profiles of the example took 1.43 minutes for compile and go. Considerable speeding-up is possible when the multiplicity of hits is not required.

Although the program was designed primarily to process Wiswesser Line-Formula Chemical Notations,[5] only a few modifications are required to adapt it to any other linear notation; the subject of the notation need not even be chemical, indeed, concordances or statistics of syllables, words, and phrases in texts of natural languages can be prepared with the unmodified program.

ACKNOWLEDGMENT

My thanks are due to Dr. A. Betser for encouragement in the work, and to the NSF for the travel grant which made my attendance at this meeting possible.

LITERATURE CITED

(1) Backus, J. W., *Proc. Intern. Conf. Inf. Proc.*, UNESCO, Paris, pp. 125–32, 1959, Oldenburg, Munich, 1960.

(2) Bonnett, H. T., and D. W. Calhoun, J. CHEM. DOC. **2**, 2–6 (1962).

(3) Brasie, W. C., and D. W. Liou, *Chem Eng. Progr.* **61** (5), 102–8 (1965).

(4) Ofer, K. D., C. N. Rice, R. B. Bourne, and S. W., Logan, "A Pilot Study for the Input to a Chemical-Structure Retrieval System," 151st Meeting, ACS, Pittsburgh, Pa., March 23, 1966.

(5) Smith E. G., W. Wiswesser, "Revised Rules of W. J. Wiswesser Line-Formula Chemical Notation," 2nd ed., Crowell Co., New York, 1967.