Use of the Double-KWIC Coordinate Indexing Technique for Chemical Line Notations*

ANTHONY E. PETRARCA,** SAULI V. LAITINEN,† and W. MICHAEL LAY††
Department of Computer and Information Science
The Ohio State University, Columbus, Ohio 43210

Received March 11, 1971

Application of the double-KWIC coordinate indexing technique to linear notations is described. The study was performed on three notations—WLN, IUPAC, and MCC—for comparative purposes. The results show that this new automatic indexing technique can be applied with the same effectiveness shown earlier for indexes derived from words in titles or title-like phrases, the basic difference being that linear notation symbols rather than words are used as the indexing units. The ease with which these indexes enable one to coordinate two or more notation symbols with each other is extremely beneficial for increasing the specificity of search when there are a large number of notations containing the symbol (or set of symbols) used as the primary access point in the index.

Despite the progress made in the development of computer search systems for identifying structures and substructures, manual searching of structural information by means of printed indexes is still valuable. Limitations concerning the use of structural diagrams and traditional chemical nomenclature for this purpose have been well documented.[1] To circumvent some of the problems, much work has been done in recent years concerning the use of linear notations for storage and retrieval of chemical structural information. These linear notations, which represent structures of chemical compounds by compact linear sequences of alphameric symbols, are not only useful for computerized handling of structural information, but can also be utilized for preparation of printed indexes which are quite useful to anyone who is knowledgeable about the notations.

The simplest form of such an index is an alphameric listing of the notations. Portions of such listings for the two notations of major interest—the Wiswesser Line Notation (WLN) and the IUPAC Notation—are shown in Figure 1. They are of limited use, however, because the rules for constructing the notation cause some structural features to be put into indexing prominence, while others may be scattered throughout the index. The indexing performance of these two notation systems has been reviewed by Bonnett.[2]

Permuted notation indexes, which are similar in principle to KWIC indexes of titles, sentences, or phrases, enable one to circumvent problems resulting from hierarchical ordering rules. Preparation of such permuted indexes for WLN's (Figure 2) has been described by Granito and coworkers,[3-5] and the use of such indexes for support of computerized structure handling systems based on both WLN[6-8] and IUPAC[9] notations has also been described. The technique has also been applied to fragments (or screens) derived from WLN[7] and the Mechanical Chemical Code (MCC) of Lefkovitz[10] (Figure 3).

Recently, we described a new type of automated index known as the Double-KWIC Coordinate Index.[11 12] To provide the necessary perspective for later discussions, construction of Double-KWIC coordinate index entries is

IUPAC	B6ZN:2/3A5ZN12(XC$_3$)3	36
	B6ZN:3/2A5ZN1C	147
	B6ZN:4/C$_2$:2EQ/N$_2$:2/2C$_3$	117
	B6ZN:4CEQ/N$_2$:2/C$_3$:3EQ;NC/B6	20
	B6ZN:4CEQ/N$_2$:2/2C$_3$	112
	B6ZN1CEQa4CQ3C6Q5	34
	B6ZN1CQ3(CENQ/2C$_2$X)4C6Q5	107
	B6ZN1CQ3CENQ4C6Q5	107
	B6ZN1CQ3C46Q5	·120
WLN	T6N CN ENJ BZ DYQR DE& FM2N2&2	71
	T6N CN ENJ BZ DYQR DG& FM2- AT6NTJ	92
	T6N CN ENJ BZ DYQR DG& FM2N2&2	70
	T6N CN ENJ BZ DYQR DO1& FMY&1N1&1	113
	T6N CN ENJ BZ DYQR DO1& FM2N2&2	75
	T6N CN ENJ BZ DYQR& FMY&1N1&1	110
	T6N CN ENJ BZ DYQR& FM2- AT6NTJ	91
	T6N CN ENJ BZ DYQR& FM2N2&2	65
	T6N CN ENJ BZ DYQR& FM3N4&4	121

Figure 1. Simple alphameric listings of IUPAC and WLN chemical notations

	↓	
L55 A CYTJ C	U1 DSWR	97749
L6TJ AO 2PS&S2	U1R	98622
1	U1UYO1	98689
FR D1XV1&VO2&2	U1	97357
/U	U2-GE-R&R&1/ &711	97478
T5SYN	V EHJ BUM CR DG BVO2	97965
T B566 BN	V EVN HN DHJ D3 L	97713
T56 B	V FVOTJ AQ CQ DQ D	97314
L67 G	V HU&TJ CO1VQ H	98750
L67 G	V JU&TJ COV1 DOV1	98557
T567/FL 2AE L BVO	V JVTJ IO1	97674
T D3 B556 BN EN J	V MVTTT&J EVR DI& GO1 H1OVZ KZ L	98017
T C666 BS CUT&TJ D	VH	97646
T56 BN DN FNVN	VJ B CNZ2OV1 F H2U1 &222	98050
L7	VJ BQ CNUNR DNW& EG	96132
T56 BVN	VJ C2Q GVQ	98584
T56 BSS	VJ	98797
T6	VM DNJ CSH E1YVQUNQ F	97370

Figure 2. Permuted (KWIC) index entries for WLN's

*Presented in part before the 5th Middle Atlantic Regional Meeting of the American Chemical Society, The University of Delaware, Newark, Del., April 1970.

†Present address, Helsinki Technological University Library, Oteniemi, Finland.

††Present address, Department of Computer Science, University of Maryland, College Park, Md.

**To whom correspondence should be addressed.

WLN		MCC	
+R MVQ	791	C/b *a	1
+A MVR	570	*a/Lc	7
+R MVR	135	Za_2SXb c	8
+R MVZ	059	*a/L c	7
+A MVZ	476	C/O c	2
+AMYM MYMA	456	C/ab	3
+A MYMMR	693	C/b*a	1
+A MYMMYMA	476	C/Oc	2
+A MYMMYZM	032	*a/ Lc	7
+A MYSNRR	312	N/ba	4
+R MYSSYSNAA	680	N/bc	5
+AMYM MYZM	032	N/Q	6
+A MYZM	259	C/ Oc	2
+RSW MZ	123	N/ Q	6
+RV MZ	234	Za_2 SXbc	8
+R MZ	468	Za_2SXbc	8

Figure 3. Permuted (KWIC) list of WLN and MCC fragments (or screens)

```
            THE NOMENCLATURE OF HIGHLY FLUORINATED MOLECULES =  TITLE

                                                    MAIN TERM
                                                    EXTRACTED
  MAIN        NOMENCLATURE
  TERM          HIGHLY FLUORINATED MOLECULES = THE * OF
                FLUORINATED MOLECULES = THE * OF HIGHLY
                MOLECULES = THE * OF HIGHLY FLUORINATED
SUBORDINATE
  TERM        MOLECULES
                NOMENCLATURE OF HIGHLY FLUORINATED * = THE
                HIGHLY FLUORINATED * = THE NOMENCLATURE OF
                FLUORINATED * = THE NOMENCLATURE OF HIGHLY

              FLUORINATED MOLECULES
                NOMENCLATURE OF HIGHLY * = THE
                HIGHLY * = THE NOMENCLATURE OF
```

Figure 4. Construction of Double KWIC Coordinate Index entries

illustrated in Figure 4, and a comparison of KWIC and Double-KWIC entries for searching under a particular concept of primary interest is illustrated in Figure 5. Thus, in contrast to the KWIC index, which allows one to gain ready access to information on the basis of a single concept at a time, the Double-KWIC coordinate index facilitates coordination of secondary concepts with the concept of primary interest. (Granito[4,5] describes the use of a QUICK-SCAN column in his permuted line notation index to facilitate coordination of secondary concepts; but, as will be shown later, it does not enable one to do so as easily as does the Double-KWIC Coordinate Index.)

In view of the preponderant number of instances where coordination of two or more functional groups is desirable for a search of chemical structural information, we decided to explore the use of the double-KWIC coordinate indexing technique on chemical line notations. The study was performed on three notations—WLN, IUPAC, and MCC—to determine if any particular notation offered any unusual advantage over the others as far as the use of this indexing technique is concerned.

SAMPLES CHOSEN FOR STUDY

A sample of about 250 WLN's was obtained from an *Index Chemicus Registry System* (*ICRS*) tape (Institute for Scientific Information, Philadelphia, Pa.) by extracting every 50th notation. (Compounds not coded in WLN were excluded from consideration.) For comparative studies a subset of about 50 notations, corresponding to every 250th notation in the original *ICRS* file, was decoded into structural formulas[13] and encoded into the corresponding IUPAC[14] and MCC[15] notations as illustrated in Figure 6. Since these last two notation systems require upper- and lower-case letters as well as other special characters not available on standard card punch equipment, appropriate substitution symbols and lower-case flags had to be used for these notations in order to perform the study. [For the IUPAC notation a period (.) preceding a capital letter was used to denote the corresponding lower-case letter, a period preceding a number to denote a subscript, and a comma (,) preceding a number to denote underlining. These flags have been used by Dyson[16] for similar purposes. Superscripts plus (+) and minus (−) were represented without flags. Also, the less than sign (<) was used as a replacement symbol for brackets. For the MCC notation the period was used as a flag for subscripts and for the lower-case letters of two-character element symbols as in the IUPAC notation, but the upper-case letters A, B, and U were used as replacement symbols for the lower-case letters a, b, and c when they represented the carbon fragments CH, CH_2, and CH_3 respectively.[17]] For the most part, the mechanics of the study are illustrated on the Wiswesser notation to maintain coherence. However, some examples from IUPAC and MCC are also included for comparative purposes.

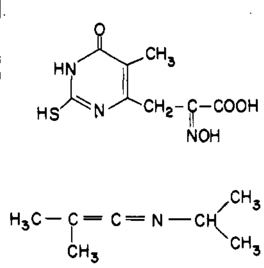
```
                              KWIC INDEX

             CARBOHYDRATE NOMENCLATURE =                                  78
 SOME PROBLEMS IN POLYMER NOMENCLATURE =                                  74
UNCTIONING OF BIOCHEMICAL NOMENCLATURE =   THE ORGANIZATION AND F         72
LEMS =           INORGANIC NOMENCLATURE IN 1966: PROGRESS AND PROB        67
CULES =                THE NOMENCLATURE OF HIGHLY FLUORINATED MOLE        82
                       THE NOMENCLATURE OF ORGANIC CHEMISTRY =            64


                       DOUBLE KWIC COORDINATE INDEX

NOMENCLATURE
  BIOCHEMICAL * = ...........THE ORGANIZATION AND FUNCTIONING OF  72
  CARBOHYDRATE * = ..............................................  78
  CHEMISTRY = ................................THE * OF ORGANIC  64
  FLUORINATED MOLECULES = ......................THE * OF HIGHLY  82
  FUNCTIONING OF BIOCHEMICAL * = ...........THE ORGANIZATION AND  72
  HIGHLY FLUORINATED MOLECULES = .......................THE * OF  82
  INORGANIC * IN 1966: PROGRESS AND PROBLEMS = .................  67
  MOLECULES = ......................THE * OF HIGHLY FLUORINATED  82
  ORGANIC CHEMISTRY = ..................................THE * OF  64
  ORGANIZATION AND FUNCTIONING OF BIOCHEMICAL * = ...........THE  72
  POLYMER * = ................................SOME PROBLEMS IN  74
  PROBLEMS = ..................INORGANIC * IN 1966: PROGRESS AND  67
  PROBLEMS IN POLYMER * = ................................SOME  74
  PROGRESS AND PROBLEMS = ..................INORGANIC * IN 1966:  67
```

Figure 5. Comparison of conventional KWIC entries and Double KWIC Coordinate Index entries for a given index term derived from the same set of titles



1 T6VM DNJ CSH E1YVQUNQ F
2a B6ZN13(C_3X3ENQ2)4C5EQ6S2
2b B6ZN13(C.3X3ENQ2)4C5EQ6S2
3a NCSHMLCcC1D7bCNQL11Q
3b NCSHMLCUC1D7BCNQL11Q

1 1Y&UCUNY
2a C_3C2E2:3EN/2C_3
2b C.3C2E2:3EN/2C.3
3a cCcC2Nacc6
3b UCUC2NAUU6

Figure 6. Chemical structures and corresponding line notations: 1-WLN, 2a-IUPAC, 2b-Modified IUPAC with flags and/or replacement symbols, 3a-MCC, 3b-Modifed MCC with flags and/or replacement symbols

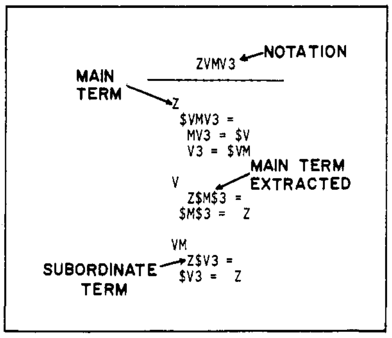


Figure 7. Construction of double-KWIC coordinate index entries from linear notations

```
                                      ↓
                                      T56 BV FVOTJ AQ CQ DQ D
                                      T56 BVNVJ C2Q GVQ
                                      T567/FL 2AE L BVOV JVTJ IO1
                                      T6N CNJ BZ DMR BVQ& FG
                                      T6NJ B2N1&1&1R CF DO2
                                      T6O COTJ BR& DYS3&S3 EO1R& F1O1R
             T56 BN DM FN HNJ IS- B T6OTJ CQ DQ EQ F1O
     6 HXO RVTJ A E G PQ QO1 TO1 H-& B T6OXTK E              T F5 E5 B66
                                      T6VM DNJ CSH E1YVQUNQ F
                                      T66 AK DYVT&J DUNQ &G
                                      T66 ANTJ E1MR BZ DXFFF
                                      T66 BN DN CN JNJ CR& EZ HVO2 IZ
                                      T66 BN DN EHJ CXGGG DR& EXGGG EO4
                                      T66 CNO EHJ BO1
```

Figure 8. KWIC index entries illustrating randomized ordering of secondary concepts (M and N) under primary concept T

CONSTRUCTION OF THE INDEX

Construction of a Double-KWIC coordinate index from linear notations is similar to construction from word strings —i.e., titles, sentences, phrases—except that the operations are performed on individual characters rather than words. The steps involved (Figure 7) are as follows:

> Each significant character or group of characters of the notation is extracted as a main term and replaced by a dollar sign ($) to indicate its position in the notation.
>
> After the first step is performed on each of the notations to be processed, the KWOC-type entries are sorted on the main terms. If the number of notations containing a particular main term is less than or equal to an arbitrarily chosen threshold value (3 for our studies), the nonpermuted notations are posted as subordinate entries. If the number of notations is greater than the threshold value, the notations are rotated so that the remaining significant symbols in each notation appear in turn in the subordinate index column as part of wrap-around subordinate entries.
>
> The resulting entries are then sorted both with regard to main terms (primary sort) and subordinate terms (secondary sort). If the extracted main term immediately precedes the subordinate index term, the dollar sign is placed in the left margin (at print time).

Significance of main terms and subordinate terms is established on the basis of appropriate edit programs, stoplists, and other methods of vocabulary control to be discussed later. The index generating programs were written in PL/1 and were executed on an IBM 360/75 computer operating under OS 360/MVT at The Ohio State University.

COMPARISON OF KWIC AND DOUBLE-KWIC

In an ordinary KWIC index of linear notations, random ordering of subordinate symbols requires examination of the entire notation to allow coordination of a second symbol with a first (except for a few combinations of symbols which are always contiguous to each other). For high occurrence symbols, the search time can become quite significant. This is illustrated on a smaller scale for a KWIC index of WLN's in Figure 8 where such a search would be required to locate all nitrogen heterocycles under the KWIC entries for the heterocyclic ring symbol T. (Granito's QUICK-SCAN area[4,5] alleviates this problem somewhat by providing an auxiliary compact string of all of the significant symbols for which index entries are created. The compactness is achieved by exclusion of locants and by listing each significant symbol only once, regardless of the number of times it occurs in the notation. Admittedly it is easier to locate a second symbol in this compact nonredundant linear string than it would be to locate it in the KWIC index entry itself, but the secondary symbols in the QUICK-SCAN area are still randomly ordered.) In the corresponding Double-KWIC index (Figure 9), all of the M and N symbols for nitrogen can easily be located in an ordered list of subordinate entries under the main index term T. False drops can easily be ascertained by examination of the surrounding context (in this case by checking to see if the M or N symbols are enclosed within the T-J ring-enclosing symbols). To keep the size of the double-KWIC coordinate index within reasonable limits, the subordinate entries are not permuted when the number of postings under the main index term is low. This treatment (Figure 10) is based on the premise that tradeoffs in the scanning movements required by the human eye to locate a secondary concept with the appropriate relationships under these circumstances favor the use of nonpermuted KWOC-type subordinate entries rather than permuted double-KWIC subordinate entries.

```
T*****************************************************************
  M CR DG BVO2 ..............................$5SYNV EHJ BU   97965 28
  M DNJ CSH E1YVQUNQ F ..............................$6V   97370  1
  M FN HNJ IS- B$6O$J CQ DQ EQ F1Q................$56 BN D   97346  9
  MR BVQ& FG..................................&6N CNJ BZ D   98134 11
  MR BZ DXFFF..................................$66 AN$J E1   96798  9
  MVOX E2/ &711....................../- D$5VNV$J BOVY4MVO1R&   97804  9
  MVO1R&MVOX E2/ &711....................../- D$5VNV$J BOVY4   97804  9
  N CNJ BZ DMR BVQ& FG..................................$6   98134 11
  N DHJ D3 L...............................$ B566 BNV EVN H   97713 14
  N DM FN HNJ IS- B$6O$J CQ DQ EQ F1Q................$56 B   97346  9
  N DN EHJ CXGGG DR& EXGGG EO4........................$66 B   98422  3
  N DN FNVNVJ B CNZ2OV1 F H2U1 &222...................$56 B   98050 36
  N DN GN JNJ CR& EZ HVO2 IZ..........................$66 B   97361 34
  N EHJ CXGGG DR& EXGGG EO4........................$66 BN D   98422  3
  N EN JV MV$$$&J EVR BI& GO1 H1OVZ KZ L........T D3 B556 B   98017 15
  N FNVNVJ B CNZ2OVI F H2U1 &222...................$56 BN D   98050 36
  N GN JNJ CR& EZ HVO2 IZ..........................$66 BN D   97361 34
  N HN DHJ D3 L..............................$ B566 BNV EV   97713 14
```

Figure 9. Double-KWIC index entries illustrating alphabetically ordered list of secondary concepts (M and N) under primary concept T

```
-AS-**************************************************************
  S$1&1&1  & 2-ZN-EE.......................................  98524 16
  T5$J A1O1 BR& ER.........................................  98391  9

-CO-**************************************************************
  SUYR&S 3$................................................  98468 15
```

Figure 10. Some main terms with nonpermuted subordinate entries

```
VQ****************************************************************
  G....................................T6N CNJ BZ DMR B$& F  98134 11
  L67 GV HU&TJ CO1$ H........................................  98750 11
  M DNJ CSH E1Y$UNQ F....................................T6V  97370  1
  MR B$& FG..............................T6N CNJ BZ D  98134 11
  N CNJ BZ DMR B$& FG....................................T6  98134 11
  NJ BZ DMR B$& FG.....................................T6N C  98314 11
  NJ CSH E1Y$UNQ F...................................T6VM D  97370  1
  NQ F.................................T6VM DNJ CSH E1Y$U  97370  1
  NVJ C2Q G$..........................................T56 BV  98584  4
  O1$ H.......................................L67 GV HU&TJ C  98750 11
  Q F..................................T6VM DNJ CSH E1Y$UN  97370  1
  Q G$..........................................T56 BVNVJ C2  98584  4
  R B$& FG......................................T6N CNJ BZ DM  98134 11
  SH E1Y$UNQ F..................................T6VM DNJ C  97370  1
  TJ CO1$ H.........................................L67 GV HU&  98750 11
  T56 BVNVJ C2Q G$.........................................  98584  4
  T6N CNJ BZ DMR B$& FG....................................  98134 11
  T6VM DNJ CSH E1Y$UNQ F...................................  97370  1
  U&TJ CO1$ H..........................................L67 GV H  98750 11
 $UNQ F.....................................T6VM DNJ CSH E1Y  97370  1
```

Figure 11. A main term for a functional group represented by contiguous WLN symbols

```
R*****************************************************************
  G BVO2...............................T5SYNV EHJ BUM C$ D  97965 28
  G FM3 DSZW..........................................ZSW$ B  97520 68
  G$ CNW F- 2..............................................WS  97134  1
  G$ DG....................................................Z2Y  98366 18
  G...................................................Z2YG$ D  98366 18
  NW F- 2...............................................WSG$ C  97134  1
  NW& EG..................................L7VJ BQ CNUN$ D  96132  6
  VO1$&Y.....................................WN$ DOVYM  98789 16
  VO1.......................................T4NTJ A1$& B  97990  3
  VO2 IZ.........................T66 BN DN GN JNJ C$& EZ H  97361 34
  VO2&2U1..........................................F$ D1XV1&  97357  1
  VO2................................T5SYNV EHJ BUM C$ DG B  97965 28
  VQ& FG.......................................T6N CNJ BZ DM$ B  98134 11
  WN$ DMMYUS&MV2YSW$&SW$..................................  98073 11
  WN$ DOVYMVO1$&Y.........................................  98789 16
  Z DM$ BVQ& FG......................................T6N CNJ B  98134 11
  Z DXFFF.....................................T66 ANTJ E1M$ B  96789  9
  Z HVO2 IZ.........................T66 BN DN GN JNJ C$& E  97361 34
  ZSW$ BG FM3 DSZW........................................  97520 68
```

Figure 13. Some examples of meaningful structural units for which the component symbols are not always contiguous to each other in the notation

```
V*****************************************************************
  M CR DG B$O2................................T5SYN$ EHJ BU  97965 28
 $M DNJ CSH E1Y$QUNQ F.....................................T6  97370  1
  M$OX E2/ &711......................../- DT5$N$TJ BO$Y4M$O1R&  97804  9
  M$R D-SI-1&1&1............................................4  97326  2
  N CNJ BZ DMR B$Q& FG.....................................T6  98134 11
 $N HN DHJ D3 L..............................T B566 BN$ E  97713 14
  N JNJ CR& EZ H$O2 IZ...........................T66 BN DN G  97361 34
  N$ E$N HN DHJ D3 L....................................T B566 B  97713 14
 $N$J B CNZ2O$1 F H2U1 &222.....................T56 BN DN FN  98050 36
  N$N$J B CNZ2O$1 F H2U1 &222....................T56 BN DN F  98050 36
  NJ BZ DMR B$Q& FG.......................................T6N C  98134 11
                                    .
                                    .
                                    .
  Z DMR B$Q& FG.......................................T6N CNJ B  98134 11
 $Z KZ L...........T D3 B556 BN EN J$ M$TTT&J E$R BI& GO1 H1O  98017 15
  Z$MNUYR&1O1XQR&R.........................................  97956  5
  Z2O$1 F H2UL &222..................   T56 BN DN FN$N$J B CN  98050 36
```

Figure 12. Double-KWIC index entries for amides. (Only the subordinate entries M, N, and Z which are preceded or followed by $ have the requisite syntactic relationships with the main term V)

```
SEQ  CT TERM                      SEQ  CT TERM
  1   5 -                          53  10 L
  2   2 -A                         54   1 L5
  3   2 -AS                        55   6 L6
  4   2 -AS-                       56   1 L7
  5   1 -C                         57  12 M
  6   1 -CO                        58   1 MM
  7   1 -CO-                       59   1 MN
  8   1 -E                         60   1 MP
  9   1 -EE                        61   3 MR
 10   1 -G                         62   4 MV
 11   1 -GE                        63   1 MVR
 12   1 -GE-                       64   1 MY
        .                          65   1 M3
        .                          66  27 N
        .
```

Figure 14. Portion of the potential-main-term list generated in the first processing phase

The present program for generating the main terms allows us to extract character strings of any desired length. This permits extraction of meaningful combinations such as VQ for carboxylic acids in the Wiswesser notation, a feature that offers definite advantages for coordinating a functional group of this type with other functional groups (see Figure 11). However, because the notation rules allow such combinations to occur in the reverse order, a significant amount of scattering might occur depending on how the viewpoint of the searcher coincides with the viewpoint from which the index is created. For example, amides can be represented by the combinations, VM, VN, VZ, and their converses in the Wiswesser notation. If one is interested in all amides, this information would be scattered in six different locations under those headings; however, the double-KWIC index that we have generated brings all of them together in an ordered list of subordinate entries under the main term V. Some representative examples are illustrated in Figure 12. Note that the amide linkages can be differentiated easily from the nonamides by the occurrence of the $ immediately before or immediately after the subordinate symbol for nitrogen (M, N, or Z).

An additional problem occurs with regard to other meaningful structural units that one might be tempted to extract as main terms for WLN's. Some cases in point are benzene derivatives such as benzoic acids, halobenzenes, etc., where the R symbol for benzene is not always contiguous to the symbol for the functional group. Some examples are illustrated in Figure 13 for chlorobenzene derivatives (G subordinate entries), nitrobenzene derivatives (NW and WN subordinate entries), benzoic acid derivatives (VQ and VO subordinate entries), and aniline derivatives (Z subordinate entries). Because of relationships such as the above, most of the main terms we extracted were one-character terms. The only exceptions were the notations for those elements requiring more than one character—e.g., -AS-, -CO-, etc.—and some selected examples of structural units such as those discussed earlier—e.g., VQ, VM, VZ, etc.

MAIN TERM SELECTION, STOP-LISTS, AND OTHER METHODS OF VOCABULARY CONTROL

With the present version of the programs, generation of double-KWIC coordinate indexes of the type described herein requires two processing phases and two human-interface steps. In the first processing phase, the human interface is needed to supply the input data and the program processing specifications to generate a list of potential main terms such as those illustrated in Figure 14. The processing program also generates sequence numbers for these terms and counts the number of notations in which each potential main term occurs. At this point, the second human interface is required for selection of the actual main terms desired in the printed index. For example, in an index generated from the potential main terms illustrated

```
PRIMARY STOP-LIST:  AHJXY&123456789-/

STOP-FLAG:  Blank (Excludes generation of main terms starting with a
                   a character preceded by a blank space.)

SUBORDINATE STOP-LIST:  AH12&-J/

OVERRIDE LIST:  AH12

EXAMPLE

     MAIN TERM GENERATION:  T66 ANTJ E1MR BZ DXFFF


     SUBORDINATE ENTRIES GENERATED UNDER  M

            T66 ANTJ E1$R BZ DXFFF
```

Figure 15. Control parameters and their effects on generation of main terms and subordinate entries in the double-KWIC coordinate index

in Figure 14, some of the actual main terms selected were #4 (-AS-), #7 (-CO-), #12 (-GE-), #53 (L), #57 (M), and others. A conventional KWIC index of the notations (generated by another program) is helpful, but not necessary, for making some of the selections. Once the selections have been made they are input via their sequence numbers together with other processing specifications required to obtain the desired output from the second processing phase.

Some of the processing specifications which have to be supplied (these may vary for different applications) will now be discussed. As mentioned earlier one can specify a range of lengths for the character strings one may wish to extract as potential main terms. This is done by specifying the lower bound and the upper bound of the desired range as input parameters. Some of the terms shown in Figure 14 resulted from our experimentation with a range of 1–4 on part of the input and a range of 1–2 on the remainder. We experimented with the 1–4 range to compare the results of listing the two-character WLN symbols as main terms both with and without their hyphen delimiters. We subsequently opted in favor of listing them without delimiters for practical reasons, one of which was that we only had to generate potential main terms of length 1–2.

To reduce the number of potential main terms generated, two other control parameters are available (Figure 15). One is a primary stop-list which will prevent generation of main terms beginning with any character on that list. The primary stop-list used for our WLN indexes is shown in Figure 15. The other control parameter, which we call a "stop-flag," enables us to define a character which will exclude generation of main terms beginning with the stop-flag and also excludes generation of terms beginning with the character immediately following the stop-flag. The stop-flag also excludes generation of permuted subordinate entries beginning with those characters as well. The use of the blank character as a stop-flag for WLN's enabled us to eliminate automatically all locants and multipliers from consideration as main terms or subordinate terms in the index. Without such a feature, the utility of both KWIC and double-KWIC indexes of WLN's would be significantly impaired owing to creation of many ambiguous index entries which would require considerable visual scanning of the context to isolate those entries having the desired meaning.

There are two additional control parameters which apply to generation of permuted subordinate entries (also illustrated in Figure 15). One of these is a subordinate stop-list which excludes generation of subordinate entries beginning with the characters on that list. The other is an override list which identifies certain symbols on the subordinate stop-list for which subordinate entries are to be generated only if the subordinate symbol immediately precedes or follows the main symbol(s) extracted from the original notation. The net effect from the use of all of these control parameters is illustrated on a sample notation in Figure 15. The resulting WLN indexes contained approximately 45 entries per notation. However, detailed analysis and evaluation of the index entries suggested that some modifications in program design (to be discussed later) coupled with some justifiable changes in stop-lists and other control parameters would easily reduce that figure to approximately 20 entries per notation and actually improve the over-all quality of the index in the process.

COMPARISON OF WLN, IUPAC, AND MCC NOTATIONS

All of the techniques described above for the WLN's were applied in similar fashion to the IUPAC and the MCC notations but, obviously, with different values for the control parameters. A portion of the double-KWIC coordinate index prepared for the IUPAC notations is illustrated in Figure 16, and a portion of the index prepared for the MCC notations is illustrated in Figure 17.

As to the effectiveness of the indexes produced from the three notations, there is little doubt about the greater ease with which one can coordinate more than one concept from

```
X*********************************************************************
  /$C.2...............A65ZN3579(C.3E2)5C379E1EQ46:8N(N)C.22:  98050 36
  /$CN)9:,1,3CEQ/B6I2.....A65.23.B17,1,1ZN7,1,3C5QCBEQ36N4(C  98017 15
  /$CNC.4(NC$/2C(C).3)3:4$/1A5ZN1C.23Q25:4/.$............B6C  97804  9
 $/C.2)2............................B6F:4/C.5(C.2EQ)2E4(C  97357  1
 $/C.2)3:4/A5ZN1ZS3EQ5EN2..........................B6C.H1(C  97965 28
 $/C.2)4:8/B6...........................B6.2ZN3679N5,1,0(C  97361 34
 $/C/B6...............................B6NO.2:4Q/C.4C3EQ1:2N/C  98789 16
 $/1A5ZN1C.23Q25:4/.$............B6C/$CNC.4(NC$/2C(C).3)3:4  97804  9
  A4ZN:2C$/C.........................................B6:C/  97990  3
  A5ZN1C.23Q25:4/.$............B6C/$CNC.4(NC$/2C(C).3)3:4$/1  97804  9
  A65.23.B17,1,1ZN7,1,3C5QC8EQ36N4(C/$CN)9:,1,3CEQ/B6I2.....  98017 15
  A65ZN3579(C.3E2)5C379E1EQ46:8N(N)C.22:/$C.2...............  98050 36
  A761.BC4E4EQ3:9Q/C.2$2....................................  98750 11
  B6.2ZN3679N5,1,Ø(c$/C.2)4:8/B6 ...........................  97361 34
  B6I2.....A65.23.B17,1,1ZN7,1,3C5QC8EQ36N4(C/$CN)9:,1,3CEQ/  98017 15
  B6ZN13C5EQ4S2(C.3$3ENQ2)6.................................  97370  1
 $C.2)9,1,Ø...................................A761.BE6EQ3(  98557  3
 $C.2...............A65ZN3579(C.3E2)5C379ELEQ46:8N(N)C.22:/  98050 36
  C$1..................................B6ZN13C.H4N2:6N/2B6  98134 11
  C$4EQ79.....................................B65ZN8(C.2Q2)8  98584  4
```

Figure 16. Some double-KWIC coordinate index entries generated from the modified IUPAC notation

```
Q*********************************************************************
  A$BA$AB$011,17..........................UR-6S5XOBBACUBBACUBB  97346 19
  A$COC$BBOCOA2,7$2.........................................UC  97314  6
 $AB.3AUB4..................................................UA  97862 10
 $AB$011,17.............................NANCMANCD4CD1SAA$BA  97346  9
 $AOBRA120A9B021.........................................RBOAA  98208 28
 $B.3ACOBAUBA5U.............................................  98338  5
  B$C6,110...........................................$LRC70NB  98584  4
  B$011,17........................NANCMANCD4CD1SAA$BA$A  97346  9
 $BA$AB$011,17................................NANCMANCD4CD1SAA  97346  9
 $BA19BBA18A14BB13.....................UR-6S5XOBBACUBBACUBBA  98276 19
 $BBA15A11BA1Ø05........UABOCBB2A5UACUBBACUAOUB18COAOUC16,24  96272 14
 $BBOCOA2,7$2...........................................UCA$COC  97314  6
 $BOBCRND18MLZ..........................................(R).2C  97956  5
  C$BBOCOA2,7$2..........................................UCA$CO  97314  6
  C$BOBCRNC18MLZ..........................................(R).2  97956  5
 $COC$BBOCOA2,7$2...........................................UCA  97314  6
 $C6,110...........................................$LRC70NBB  98584  4
  L$.............................................RB6BACUC1004B  98750 11
 $LRC70NBB$C6,110...........................................  98584  4
 $LRM8CACGNCZN1Ø............................................  98134 11
```

Figure 17. Some double-KWIC coordinate index entries generated from the modified MCC notation

an index of this type for any one of the notations. However, the usefulness of the MCC for such indexes seems to be limited by its inability to provide any readily identifiable index entries on ring systems as do the Wiswesser and IUPAC notations. Also, the larger character set required by the IUPAC notation, while seemingly a handicap at present because of current computer hardware and software constraints, may offer definite advantages when these constraints no longer exist because fewer context-sensitive uses of each symbol in the notation will be required as compared to the other notations. These are the only meaningful comparisons that can be made on the basis of the studies we have performed thus far.

CONCLUSIONS

The double-KWIC coordinate indexing technique has been shown to be applicable to chemical line notations with the same effectiveness as observed earlier for word indexes derived from titles. Our study has shown also, however, that the symbols for these notations require more syntactic analysis than we had anticipated. Although our present programs for generating double-KWIC indexes of these notations handle a limited number of syntactic relationships—e.g., the "stop-flag" treatment described earlier—a closer look at the notations suggests that other syntactic relationships can be handled with little additional effort. For example, two-character symbols such as -GE- in the Wiswesser notation can easily be identified by the hyphen immediately preceding and following the symbol, and appropriate algorithms can be developed to recognize such relationships. It should be possible also to differentiate between the heterocyclic T symbol and the T used for ring saturation, and between the use of numbers for ring sizes and for alkyl chains. Finally, it should be possible to generate an authority list that could be used to post subordinate entries under a preferred term to eliminate scattering between two headings such as VQ and QV, etc., as discussed earlier. These are some of the activities we hope to pursue to improve the quality of these indexes.

We feel that double-KWIC coordinate indexes of chemical line notations can provide a greater degree of accessibility to the type of structural information that can be derived from such notations. Consequently, they should represent a welcome addition to the tools available for retrieval of such information by organizations which have chemical structural information stored in linear notation form.

ACKNOWLEDGMENT

We thank the Institute for Scientific Information for making one of their ICRS types available to us for experimental purposes. We also thank the Office of Education for a fellowship to W. M. L. and The Ohio State University Instructional and Research Computer Center for providing much of the computer time required to perform this research. In addition, S. V. L. thanks the Finnish National Council of Scientific and Technical Information for permitting him to work at Chemical Abstracts Service under the exchange program for Visiting Information Scientists set up by the American Chemical Society; he also thanks Chemical Abstracts Service for allowing him to take some courses in computer and information science at The Ohio State University during his internship (this paper resulted from work which he performed for one of those courses). Partial support for this work by a grant (GN-534.1) from the National Science Foundation, Office of Science Information Services, is also gratefully acknowledged.

LITERATURE CITED

(1) "Chemical Structure Information Handling. A Review of the Literature 1962–1968," National Academy of Sciences, Publ. No. **1733**, Washington, D. C., 1969.
(2) Bonnett, H. T., "Chemical Notations—A Brief Review," *J. Chem. Doc.* **3**, 235 (1963).
(3) Sorter, P. F., Granito, C. E., Gilmer, J. C., Gelberg, Alan, and Metcalf, E. A., "Rapid Structure Searches via Permuted Chemical Line-Notations," *J. Chem. Doc.* **4**, 56 (1964).
(4) Granito, C. E., Gelberg, Alan, Schultz, J. E., Gibson, G. W., and Metcalf, E. A., "Rapid Structure Searches via Permuted Chemical Line Notations. II. A Key-Punch Procedure for the Generation of an Index for a Small File," *J. Chem. Doc.* **5**, 52 (1965).
(5) Granito, C. E., Schultz, J. E., Gibson, G. W., Gelberg, Alan, Williams, R. J., and Metcalf, E. A., "Rapid Structure Searches via Permuted Chemical Line Notations. III. A Computer-Produced Index," *J. Chem. Doc.* **5**, 229 (1965).
(6) Bowman, C. M., Landee, F. A., Lee, N. W., Reslock, M. H., and Smith, B. P., "A Chemically Oriented Information Storage and Retrieval System. III. Searching a Wiswesser Line Notation File," *J. Chem. Doc.* **10**, 50 (1970).
(7) Thomson, L. H., Hyde, E., and Matthews, F. W., "Organic Search and Display Using a Connectivity Matrix Derived from Wiswesser Notation," *J. Chem. Doc.* **7**, 204 (1967).
(8) Ofer, K. D., "A Computer Program to Index or Search Linear Notations," *J. Chem. Doc.* **8**, 128 (1968).
(9) Dammers, H. F., and Polton, D. J., "Use of the IUPAC Notation in Computer Processing of Information on Chemical Structures," *J. Chem. Doc.* **8**, 150 (1968).
(10) Lefkovitz, David, "Substructure Search in the MCC System," *J. Chem. Doc.* **8**, 166 (1968).
(11) Petrarca, A. E., and Lay, W. M., "The Double-KWIC Coordinate Index. A New Approach for Preparation of High-Quality Printed Indexes by Automatic Indexing Techniques," *J. Chem. Doc.* **9**, 256 (1969).
(12) Petrarca, A. E., and Lay, W. M., "The Double KWIC Coordinate Index. II. Use of an Automatically Generated Authority List to Eliminate Scattering Caused by Some Singular and Plural Main Index Terms," *ASIS Proceedings* **6**, 277 (1969).
(13) Smith, E. G., "The Wiswesser Line-Formula Chemical Notation," McGraw-Hill, New York, 1968.
(14) "Rules for I.U.P.A.C. Notation for Organic Compounds," Wiley, New York, 1961.
(15) Lefkovitz, David, "A Chemical Notation and Code for Computer Manipulation," *J. Chem. Doc.* **7**, 186 (1967).
(16) Dyson, G. M., and Riley, E. F., "Mechanical Storage and Retrieval of Organic Chemical Data," *Chem. Eng. News* **39**, 72 (1961).
(17) Lefkovitz, David, "Use of the MCC Topological Screen System," Presented at the Tutorial in Available Computer Programs for Information Retrieval. Division of Chemical Literature, 158th Meeting, ACS, New York, N. Y., September 11, 1969; also, Lefkovitz, David, and Gennaro, A. R., "A Utility Analysis for the MCC Topological Screen," *J. Chem. Doc.* **10**, 86 (1970).