Computer Storage and Retrieval of Generic Chemical Structures in Patents. 2. GENSAL, a Formal Language for the Description of Generic Chemical Structures

JOHN M. BARNARD, MICHAEL F. LYNCH,* and STEPHEN M. WELFORD

Postgraduate School of Librarianship and Information Science, University of Sheffield, Western Bank, Sheffield S10 2TN, United Kingdom

Received March 25, 1981

A formal language, GENSAL, is described which is designed for the concise and unambiguous representation of generic structures from chemical patents ("Markush formulas") in a manner which is intelligible to a chemist, yet sufficiently formalized for automatic analysis by a computer. GENSAL contains a number of facilities for showing the alternatives possible in a generic structure and any special restrictions placed on their variety. Experience to date suggests that GENSAL is capable of dealing with most of the types of generic statement encountered in patents.

INTRODUCTION

The first paper in this series[1] has outlined the reasons for the development of a special input notation, or language, for the description of generic chemical structures in patents. Such a language must allow the user to communicate effectively with a computer system for handling generic structures, without his needing a detailed understanding of the internal representation used for the structure or of the methods used for searching files of such structures. The notation required will represent a generic or "Markush" structure in a form which is intelligible to the chemist or information scientist but yet can be used to generate automatically an internal machine representation of the structure, which may then be used for searching. Since the internal representation is to be unambiguous (that is, it is to describe exactly which specific structures are covered by the generic structure, at least in as far as the patent specification itself does so), it follows that the input notation must also be unambiguous, and it is envisaged that the notation should be used for the input of both file structures and queries and possibly also for the output of the results of searches.

The representation has been termed "GENSAL" (GENeric Structure LAnguage). It is a formal language,[2] analogous to a computer programming language, and in its design some attention has been paid to the structure of modern programming languages, in particular Pascal.[3,4] The definition of a formal language may be made by using a "4-tuple"[2] as follows:

(1) a set of "terminal symbols", the words and symbols that actually appear in a GENSAL sentence;

(2) a set of "nonterminal symbols", the descriptive terms or "metasymbols" used to refer to elements of the sentence;

(3) a set of "production rules", the rules of the grammar of the language;

(4) a "sentence symbol", a member of the set of nonterminal symbols, which is the term referring to a complete sentence of the language.

In GENSAL the terminal symbols consist of various "delimiter words", such as BEGIN, IF, and THEN, punctuation and other "delimiter symbols", such as ";", "=", "D=", ",", and "<", the digits 0-9, and various data elements, such as structure diagrams and nomenclatural terms. The nonterminal symbols are the headings on the syntax diagrams shown in Figure 14, and these syntax diagrams represent the production rules for GENSAL. Throughout the text of this paper, the nonterminal symbols are shown in *italics*. Each syntax diagram is like a flow chart and shows the sequences of symbols permissible in GENSAL sentences with the order in which they may occur and the alternatives available at each point. Thus, for example, syntax diagram 7 for *parameter* shows that a *parameter* can consist either of a parameter identifier or of a single quote mark followed by either a chemical symbol or a *substituent* (defined in syntax diagram 3) and another single quote mark. One GENSAL sentence is defined as the representation of one generic structure. The syntax diagrams contain both terminal and nonterminal symbols and are arranged so that as far as possible each is defined in terms of itself (recursively) and in terms of those that precede it. The last syntax diagram, that for *structure description* is thus defined in terms of all the others, and *structure description* is therefore the sentence symbol of GENSAL.

GENERIC STRUCTURE DESCRIPTION USING GENSAL

GENSAL employs Geivandov's concept[5] of a generic structure as consisting of a (possibly vestigial) constant part to which there are attached variable parts, called *substituents*, which can vary in nature or position of attachment to the constant structure. (This use of the word "substituent" is rather wider than that normally employed by chemists and encompasses, for example, di- and polyvalent as well as monovalent radicals.) Any of the constant or variable parts can occur a constant or variable number of times, this number being indicated by a *multiplier*. In a generic structure as shown in a patent specification or abstract, a great variety of symbols is used to represent *substituents* and *multipliers*, including upper- and lower-case letters, subscripts, superscripts, etc. GENSAL standardizes these to R1, R2, R3, etc., for *substituents*, and M1, M2, M3, etc., for *multipliers*, as shown in syntax diagrams 3 and 4. During the course of a GENSAL sentence *substituents* and *multipliers* are introduced (declared), normally by appearing in a structure diagram, and are later given values (defined) in terms of chemical nature and position for *substituents* and *integer ranges* for *multipliers*. Special restrictions on the variety of the possible alternatives may then be made.

The definition of *substituents* and *multipliers* takes place in *assignment statements*, which contain facilities for assigning the same set of alternatives to groups of *substituents* or *multipliers* simultaneously, with both independent and nonindependent selection of the alternative values, or for assigning to *substituent combinations* (forming, for example, an extra fused ring). The *substituent value* may be given in several different ways, and there is scope to indicate the position at which the *substituent* is attached and any further substitution on it, down to any level of nesting. Restrictions on the variety of alternative specific structures covered by the generic structure are indicated by IF and RESTRICT *statements*. The former allow the use of one of two alternative subordinate *statements* according to whether a *condition* involving the *substituents* and *multipliers* already defined is TRUE or FALSE. The latter impose such *conditions* on the alternatives

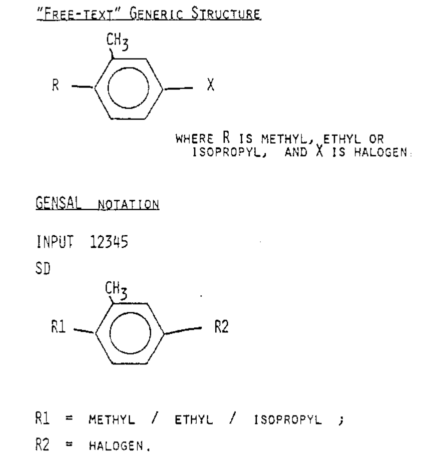

"FREE-TEXT" GENERIC STRUCTURE
CH3
R
X
WHERE R IS METHYL, ETHYL OR ISOPROPYL, AND X IS HALOGEN
GENSAL NOTATION
INPUT 12345
SD
CH3
R1
R2
R1 = METHYL / ETHYL / ISOPROPYL ;
R2 = HALOGEN.

Figure 1.

given in earlier *assignment statements*, allowing only those combinations of alternatives that result in the *condition* being TRUE.

Syntax diagram 18 shows that the overall description of a structure has an introductory heading part, consisting of a reference number, and a structure diagram for the constant part of the structure, which is followed by a series of *statements*, separated by semicolons; the sentence ends with a period. Figure 1 shows a simple generic structure and its GENSAL representation which, as can be seen, remains readily intelligible to a chemist.

As a formal language, GENSAL is designed as a means of inputting generic structures to a computer; it may equally well, however, be used as a means of describing manually such a structure in a clear and concise form, just as Algol is used for describing algorithms in a noncomputer context. It is intended that GENSAL input to a computer should be interactive, and thus the program that will read and process the GENSAL input is more correctly described as an interpreter than as a compiler, which is the analogous program for a programming language. Nevertheless, the GENSAL interpreter program, currently being developed here, is based on many of the same principles as a compiler.[6,7]

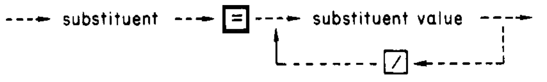
Work is still in progress on the form of the main internal representation of the generic structure, which will be derived automatically from the GENSAL input and used for fragment generation and search routines. For various reasons, given in the first paper in this series,[1] it has been decided to base this on connection tables, with partial connection tables representing the constant part of the structure and the alternative values for each *substituent* and a table at the end indicating the logical relationships involved. Homologous series terms (described below) will be represented by *parameter* lists referring to the chemical grammars[8], and "other terms", such as property-defined terms, will be included as text. The means of representing *conditions* is still under active consideration.

Most of the remainder of this paper is concerned with a formal description of the language, which will allow a full understanding of the GENSAL notations for the actual patent examples shown with the abstracts of the original specifications prepared by Derwent Publications Ltd. in Figures 2–11. A more comprehensive description of GENSAL, with further examples, is in preparation.[9]

FORMAL DESCRIPTION OF GENSAL

Structure Diagram Input. When GENSAL is being input to a computer some means is needed to enter the structure diagrams which form an integral part of the language. Any suitable chemical structure graphics system might be used, with a routine to convert its output into the connection table format required by a generic structure system. In the present work, a modification of the program developed by Feldmann and others,[10] and used in the CSSR[11] and NIH/EPA[12] substructure search systems, is being used. This has the advantage that it uses standard lineprinter characters in its display routines, and thus does not require any special hardware.

Simple Assignment Statements. The simplest form of *assignment statement* for *substituents* can be represented as follows:



(This is a simplified version of the relevant syntax diagrams.) It allows a single *substituent* on the left-hand side to be defined as having one of the values separated by the / delimiters on the right-hand side. Each alternative value is given in terms of syntax diagram 8, which contains five different possible paths:

?. This represents a completely unknown *substituent*; no information whatever is given about its nature.

Structure Diagram. This allows the *substituent* to be defined in terms of a structure diagram, which is input in exactly the same way as the main structure diagram for the constant part of the structure, and may, of course, include further *substituents*, which are declared within it.

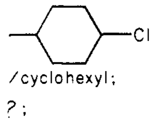
Homologous Series Identifier. This is dealt with below.

Specific Nomenclatural Term. This allows the use of a nomenclatural term to define a *substituent*. The term must represent a specific substructure, such as methyl, chloro, or phenyl, but may be a linear formula like NO2, Cl, or CN, which represents a specific substructure. Essentially this is a shorthand method of inputting a structure diagram: an operational system might well have more sophisticated routines for nomenclature translation and linear formula analysis, though development of these does not form part of our current work. At a simpler level, when the GENSAL sentence is being interpreted by the computer, a table will be searched for a record of the structure of, e.g., phenyl, and if no entry is found a suitable message will be printed at the terminal to request the input of the full structure diagram.

Other Term. This is a verbal expression that does not correspond to a specific structure, and the computer will not attempt to find a record of the structure. Examples of such terms are "easily hydrolyzed group", "nonchromophoric cation", and "electron-withdrawing group".

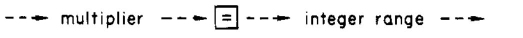
The following are simple examples of *assignment statements*:

```
R1 = methyl/ethyl/cyclohexyl;
R2 = 'electron-withdrawing group';
R3 = SD
```

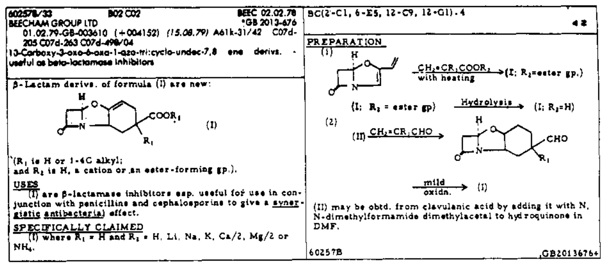


```
     /cyclohexyl;
R4 = ?;
```

and further examples may be found in the GENSAL notations for patent examples shown in Figures 2 and 3. Simple *multiplier* assignments are of the form



and enable *multipliers* to be assigned a range of *integer* values. Such a range is defined by using the *integer range* given in syntax diagram 2 and may consist of a single *integer*, or a group of *integers*, enclosed between the delimiters ⟨ and ⟩. The comma delimiter is used to separate each range fragment, which can be a single *integer* or a group of *integers* specified


60257B/33 B02 C02 BEEC 02.02.78
BEECHAM GROUP LTD *GB 2013-676
01.02.79-GB-003610 (+004152) *(15.08.79)* A61k-31/42 C07d-205 C07d-263 C07d-498/04
13-Carboxy-3-oxo-6-oxa-1-aza-tri:cyclo-undec-7,8 ene derivs. - useful as beta-lactamase inhibitors

BC(2-C1, 6-E5, 12-C9, 12-G1).4 4 2

β-Lactam derivs. of formula (I) are new:

(I)

(R₁ is H or 1-4C alkyl;
and R₂ is H, a cation or an ester-forming gp.).

USES
(I) are β-lactamase inhibitors esp. useful for use in conjunction with penicillins and cephalosporins to give a synergistic antibacterial effect.

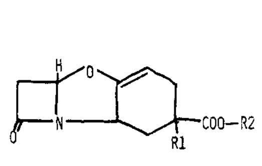
SPECIFICALLY CLAIMED
(I) where R₁ = H and R₂ = H, Li, Na, K, Ca/2, Mg/2 or NH₄.

PREPARATION
(1) → (I; R₂=ester gp.), $CH_2{=}CR_1COOR_2$ with heating
(I; R₂ = ester gp) $\xrightarrow{\text{Hydrolysis}}$ (I; R₂=H)
(2)
(II) $\xrightarrow{CH_2{=}CR_1CHO}$
$\xrightarrow[\text{oxidn.}]{\text{mild}}$ (I)

(II) may be obtd. from clavulanic acid by adding it with N,N-dimethylformamide dimethylacetal to hydroquinone in DMF.

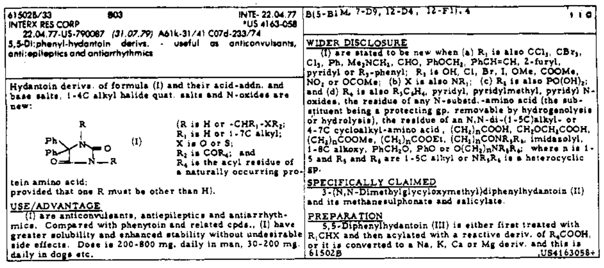
60257B GB2013676+

```
INPUT 2013676

SD
```



```
R1 = H / ALKYL <1-4> ;

R2 = H / 'CATION' / 'ESTER-FORMING GROUP' .
```

Figure 2.


61502B/33 B03 INTE- 22.04.77
INTERX RES CORP *US 4163-058
22.04.77-US-790087 *(31.07.79)* A61k-31/41 C07d-233/74
5,5-Di:phenyl-hydantoin derivs. - useful as anticonvulsants, anti:epileptics and antiarrhythmics

B(5-B1M, 7-D9, 12-D4, 12-F1).4 1 1 0

Hydantoin derivs. of formula (I) and their acid-addn. and base salts, 1-4C alkyl halide quat. salts and N-oxides are new:

(I) (R is H or -CHR₁-XR₂;
R₁ is H or 1-7C alkyl;
X is O or S;
R₂ is COR₄; and
R₄ is the acyl residue of a naturally occurring protein amino acid;
provided that one R must be other than H).

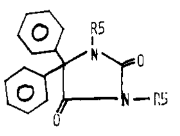
USE/ADVANTAGE
(I) are anticonvulsants, antiepileptics and antiarrhythmics. Compared with phenytoin and related cpds., (I) have greater solubility and enhanced stability without undesirable side effects. Dose is 200-800 mg. daily in man, 30-200 mg. daily in dogs etc.

WIDER DISCLOSURE
(I) are stated to be new when (a) R₁ is also CCl₃, CBr₃, CI₃, Ph, Me₂NCH₂, CHO, PhOCH₂, PhCH=CH, 2-furyl, pyridyl or R₃-phenyl; R₃ is OH, Cl, Br, I, OMe, COOMe, NO₂ or OCOMe; (b) X is also NR₁; (c) R₂ is also PO(OH)₂; and (d) R₄ is also R₃C₆H₄, pyridyl, pyridylmethyl, pyridyl N-oxides, the residue of any N-substd.-amino acid (the substituent being a protecting gp. removable by hydrogenolysis or hydrolysis), the residue of an N,N-di-(1-5C)alkyl- or 4-7C cycloalkyl-amino acid, $(CH_2)_nCOOH$, CH_2OCH_2COOH, $(CH_2)_nCOOMe$, $(CH_2)_nCOOEt$, $(CH_2)_nCONR_5R_6$, imidazolyl, 1-8C alkoxy, $PhCH_2O$, PhO or $O(CH_2)_nNR_5R_6$; where n is 1-5 and R₅ and R₆ are 1-5C alkyl or NR₅R₆ is a heterocyclic gp.

SPECIFICALLY CLAIMED
3-(N,N-Dimethylglycyloxymethyl)diphenylhydantoin (II) and its methanesulphonate and salicylate.

PREPARATION
5,5-Diphenylhydantoin (III) is either first treated with R₁CHX and then acylated with a reactive deriv. of R₄COOH, or it is converted to a Na, K, Ca or Mg deriv. and this is
61502B US4163058+

```
INPUT 4163058

SD
```



```
R5 = H / SD
         -CH— R3— R2
          |
          R1
                          ;
R1 = H / ALKYL <1-7> ;
R2 = SD
         -C — R4
          ||
          O               ;
R3 = SD   -O-      /   SD    -S-    ;
R4 = 'ACYL RESIDUE OF NATURALLY OCCURRING
         PROTEIN AMINO ACID' ;
RESTRICT  1  R5 <> H .
```

Figure 3.

by its bounds (which are shown separated by a hyphen).

If a hyphen appears immediately after the ⟨, then all *integers* from zero up to the bound are included, and if one appears immediately before the ⟩, then all *integers* from the bound upward are included.

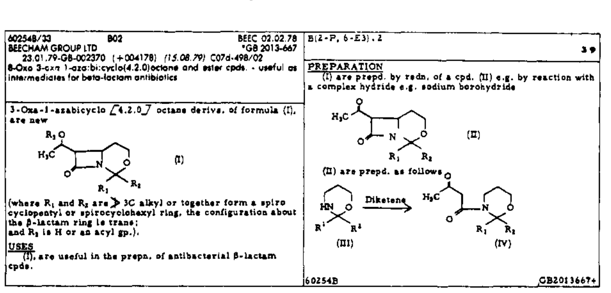
Thus the *integer range* ⟨-6,8,12,15–19,23–25,31,43-⟩ includes the *integers* 0, 1, 2, 3, 4, 5, 6, 8, 12, 15, 16, 17, 18, 19, 23, 24, 25, 31, 43, 44, 45, 46, etc., potentially up to infinity.

As can be seen from syntax diagram 1, negative *integers* are not allowed in GENSAL.

More Complex Assignments. The full syntax diagram for *assignment statement* (No. 14) and those with which it is defined (No. 10 and 11 for *substituent group* and *multiplier group*, respectively) allow much more complicated assignments to be concisely represented. For example, two *substituents* can be combined, forming perhaps an extra fused ring:

R1 + R2 = cyclopentyl/cyclohexyl

i.e., R1 and R2 combine together, forming either a cyclopentyl


60254B/33 B02 BEEC 02.02.78
BEECHAM GROUP LTD *GB 2013-667
23.01.79-GB-002370 (+004178) *(15.08.79)* C07d-498/02
8-Oxo 3-oxa 1-aza:bi:cyclo(4.2.0)octane and ester cpds. - useful as intermediates for beta-lactam antibiotics

B(2-P, 6-E3).2 3 9

3-Oxa-1-azabicyclo [4.2.0] octane derivs. of formula (I), are new

(I)

(where R₁ and R₂ are ≯ 3C alkyl or together form a spiro cyclopentyl or spirocyclohexyl ring, the configuration about the β-lactam ring is trans;
and R₃ is H or an acyl gp.).

USES
(I), are useful in the prepn. of antibacterial β-lactam cpds.

PREPARATION
(I) are prepd. by redn. of a cpd. (II) e.g. by reaction with a complex hydride e.g. sodium borohydride

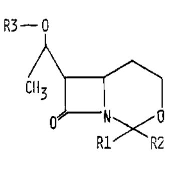
(II)

(II) are prepd. as follows

(III) $\xrightarrow{\text{Diketene}}$ (IV)

60254B GB2013667+

```
INPUT 2013667

SD
```



```
R <1-2> = ALKYL <1-3> ;
R1 + R2  =  CYCLOPENTYL / CYCLOHEXYL ;
R3  =  H  /   ACYL  .
```

Figure 4.

or a cyclohexyl ring. In this case the nomenclatural term describes the structure formed by the combination of those atoms added and those that were already present in the constant part of the structure. Figure 4 shows an example of this type of combination in a patent; in this case the possibility of combining R1 and R2 is alternative to their being separate monovalent radicals.

Similarly *multipliers* can be combined arithmetically:

M1 + M2 = ⟨5-9⟩

i.e., the sum of M1 and M2 is in the range 5–9 inclusive.

M4 − M3 = 2

i.e., M4 is 2 greater than M3.

Additionally, several *substituents* or *multipliers* can be defined simultaneously:

R⟨1-2⟩ = phenyl/cyclohexyl/cyclopentyl

i.e., R1 and R2 are both defined by the three alternatives shown. (Note that in this case the three nomenclatural terms are being used to represent monovalent radicals, substituted on the constant part of the structure, in contrast to their use above with a *substituent combination.*)

The five available assignment operators shown in syntax diagram 14 have different meanings, which may be useful when several *substituents* or *multipliers* are being defined together in a "group assignment", though in most cases encountered in patents the "=" operator is the appropriate one.

=. The *substituents* or *multipliers* in the group are independently selected from the alternatives in the *substituent definition* or *integer range* on the right-hand side of the *assignment statement.*

S=. All the *substituents* or *multipliers* in the group must have the same value, which is selected from the *substituent definition* or *integer range* on the right-hand side of the *assignment statement.*

D=. Each of the *substituents* or *multipliers* in the group must have a different value, all the values being selected from those on the right-hand side of the *statement.*

$=. Not all the *substituents* or *multipliers* in the group may be the same (which, using =, they could be), but they need not all be different.

#=. Not all the *substituents* or *multipliers* in the group may be different (i.e., at least two must be the same), but they need not all be the same.

Examples of such simultaneous assignments are as follows:

(a) R⟨1-3⟩ = phenyl/cyclohexyl/cyclopentyl;. R1, R2, and R3 can be independently either phenyl, cyclohexyl, or cyclopentyl. There are thus 27 (3 × 3 × 3) possible permutations, assuming that there are no symmetry considerations involved.

(b) R⟨4–6⟩ D = phenyl/cyclohexyl/cyclopentyl;. R4, R5, and R6 must be different, each being selected from the possibilities phenyl, cyclohexyl, and cyclopentyl. There are thus 6 (3 × 2 × 1) possible permutations.

(c) R⟨7,9–10⟩ S = phenyl/cyclohexyl/cyclopentyl;. R7, R9, and R10 must be the same. There are only three possible permutations (all phenyl, all cyclohexyl, or all cyclopentyl).

(d) R⟨11,13,15⟩ $= phenyl/cyclohexyl/cyclopentyl;. R11, R13, and R15 are not all the same, each being otherwise selected from the possibilities given. This leaves 24 possible permutations, there being three ways in which all can be the same.

(e) R⟨16–18⟩ #= phenyl/cyclohexyl/cyclopentyl;. R16, R17, and R18 are not all different, but are otherwise selected from the available possibilities. Here there are 21 possible permutations, as there are six ways in which the three may be all different.

A group *assignment statement* can begin with an *integer range*, called a "selector" which allows just some of the *substituents* or *multipliers* in the group to be assigned values from the *substituent definition* or *integer range* on the right-hand side of the *assignment statement*. For example,

⟨2–3⟩ R⟨1–5⟩ = phenyl/cyclohexyl/cyclopentyl;

means that two or three of the group of *substituents* R1, R2, R3, R4, and R5 are independently selected from the list phenyl, cyclohexyl, and cyclopentyl, the others remaining undefined at this point in the GENSAL sentence.

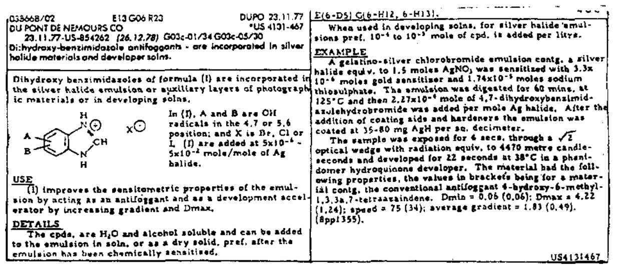
Homologous Series Identifiers and Grammars. Certain terms used in generic structures cover a range of specific substructures, all of which are alternative to each other at that point. The most common example is the term "alkyl" which covers all rooted acyclic substructures containing carbon and hydrogen only, without any unsaturations. Chemical grammars are being developed to deal with this type of term and are described in the next paper in this series.[8] Not only are they applicable to terms such as alkyl and alkenyl, which are commonly understood by chemists as "homologous series terms", but they may also be applied to many less precise terms which are none the less "structurally recognizable"—that is, terms which encompass a range of substructures that have a particular structural feature in common, such as "aryl" and "heterocyclic". Each valid homologous series identifier is associated with a list of *parameters* to the chemical grammars, each *parameter* being defined by means of an *integer range*.

As can be seen from the syntax diagram for *parameter* (No. 7), the *parameter* may be indicated by a parameter identifier, chemical element symbol, or *substituent*. The standard parameter identifiers so far being used are shown with their meanings in Figure 12, though it is likely that some additional parameters may be introduced as the chemical grammars are further developed. Thus for the homologous series identifier "alkyl", all the *parameters* will be zero, except for C, T, Q, and P, which can take on any (mutually consistent) value.

Syntax diagram 8 allows a *parameter* list to follow a homologous series identifier, thus more closely defining any of the *parameters*. This results in expressions like

alkyl C⟨3–8⟩ T⟨1–2⟩

which indicates alkyl groups containing between three and eight carbon atoms, with one or two ternary branching atoms. The *parameters* may appear in any order, or be absent altogether, in which case their default values will be the widest *integer range* compatible with those *parameters* that are present (including any implicit in the homologous series identifier itself). In view of the way in which the parameter identifier C (for carbon count) occurs almost every time a homologous series identifier is used, a shorthand has been introduced whereby the C may be omitted, provided this is



035668/02 E13 G06 R23 DUPO 23.11.77
DU PONT DE NEMOURS CO *US 4131-467
23.11.77-US-854262 *(26.12.78)* G03c-01/34 G03c-05/30
Di:hydroxy-benzimidazole antifoggants - are incorporated in silver halide materials and developer solns.

Dihydroxy benzimidazoles of formula (I) are incorporated in the silver halide emulsion or auxiliary layers of photographic materials or in developing solns.

In (I), A and B are OH radicals in the 4,7 or 5,6 position; and X is Br, Cl or I. (I) are added at 5×10^{-6} - 5×10^{-2} mole/mole of Ag halide.

USE

(I) improves the sensitometric properties of the emulsion by acting as an antifoggant and as a development accelerator by increasing gradient and Dmax.

DETAILS

The cpds. are H_2O and alcohol soluble and can be added to the emulsion in soln. or as a dry solid, pref. after the emulsion has been chemically sensitised.

E(6-D5) G(6-H12, 6-H13).

When used in developing solns. for silver halide emulsions pref. 10^{-4} to 10^{-3} mole of cpd. is added per litre.

EXAMPLE

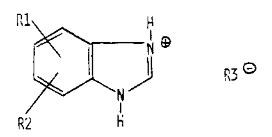
A gelatino-silver chlorobromide emulsion contg. a silver halide equiv. to 1.5 moles $AgNO_3$ was sensitized with 3.3×10^{-6} moles gold sensitiser and 1.74×10^{-5} moles sodium thiosulphate. The emulsion was digested for 60 mins. at 125°C and then 2.27×10^{-4} mole of 4,7-dihydroxybenzimidazolehydrobromide was added per mole Ag halide. After the addition of coating aids and hardeners the emulsion was coated at 35-80 mg AgH per sq. decimeter.

The sample was exposed for 4 secs. through a $\sqrt{2}$ optical wedge with radiation equiv. to 4470 metre candle-seconds and developed for 22 seconds at 38°C in a phenidomer hydroquinone developer. The material had the following properties, the values in brackets being for a material contg. the conventional antifoggant 4-hydroxy-6-methyl-1,3,3a,7-tetraazaindene. Dmin = 0.06 (0.06); Dmax = 4.22 (1.24); speed = 75 (34); average gradient = 1.83 (0.49). (8pp1355).

US4131467

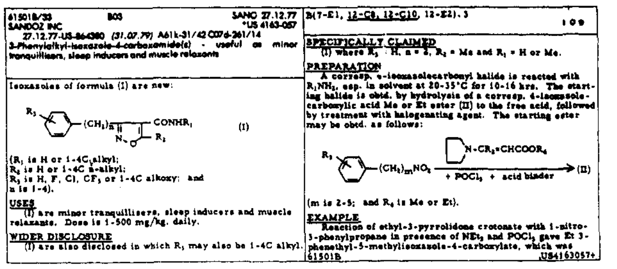
INPUT 4131467

SD



```
R1 = [<4-5>] OH;
IF R1 = [4] THEN R2 = [7] OH
        ELSE R2 = [6] OH ;
R3 = BR / CL / I .
```

Figure 5.



615018/33 B03 SANO 27.12.77
SANDOZ INC *US 4163-057
27.12.77-US-864380 *(31.07.79)* A61k-31/42 C07d-261/14
3-Phenylalkyl-isoxazole-4-carboxamide(s) - useful as minor tranquillisers, sleep inducers and muscle relaxants

Isoxazoles of formula (I) are new:

(R_1 is H or 1-4C alkyl;
R_2 is H or 1-4C n-alkyl;
R_3 is H, F, Cl, CF_3 or 1-4C alkoxy; and
n is 1-4).

USES

(I) are minor tranquillisers, sleep inducers and muscle relaxants. Dose is 1-500 mg/kg. daily.

WIDER DISCLOSURE

(I) are also disclosed in which R_1 may also be 1-4C alkyl.

B(7-E1, 12-C8, 12-C10, 12-E2). 3 109

SPECIFICALLY CLAIMED

(I) where R_3 = H, n = 2, R_2 = Me and R_1 = H or Me.

PREPARATION

A corresp. 4-isoxazolecarbonyl halide is reacted with R_1NH_2, esp. in solvent at 20-35°C for 10-16 hrs. The starting halide is obtd. by hydrolysis of a corresp. 4-isoxazolecarboxylic acid Me or Et ester (II) to the free acid, followed by treatment with halogenating agent. The starting ester may be obtd. as follows:

(m is 2-5; and R_4 is Me or Et).

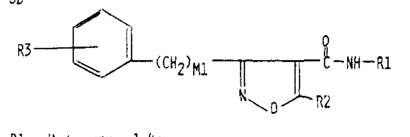
EXAMPLE

Reaction of ethyl-3-pyrrolidone crotonate with 1-nitro-3-phenylpropane in presence of NEt_3 and $POCl_3$ gave Et 3-phenethyl-5-methylisoxazole-4-carboxylate, which was

61501B US4163057+

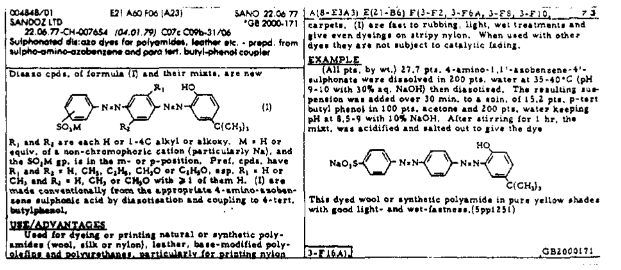
INPUT 4163057

SD



```
R1 = H / ALKYL <1-4> ;
R2 = H / ALKYL <1-4> T 0 Q 0 ;
R3 = H / F / CL / CF3 / ALKOXY <1-4> :
M1 = <1-4> .
```

Figure 6.



004848/01 E21 A60 F06 (A23) SANO 22.06 77
SANDOZ LTD *GB 2000-171
22.06.77-CH-007654 *(04.01.79)* C07c C09b-31/06
Sulphonated dis:azo dyes for polyamides, leather etc. - prepd. from sulpho-amino-azobenzene and para tert. butyl-phenol coupler

Disazo cpds. of formula (I) and their mixts. are new

R_1 and R_2 are each H or 1-4C alkyl or alkoxy. M = H or equiv. of a non-chromophoric cation (particularly Na), and the SO_3M gp. is in the m- or p-position. Pref. cpds. have R_1 and R_2 = H, CH_3, C_2H_5, CH_3O or C_2H_5O, esp. R_1 = H or CH_3 and R_2 = H, CH_3 or CH_3O with ≥1 of them H. (I) are made conventionally from the appropriate 4-amino-azobenzene sulphonic acid by diazotisation and coupling to 4-tert. butylphenol.

USE/ADVANTAGES

Used for dyeing or printing natural or synthetic polyamides (wool, silk or nylon), leather, base-modified polyolefins and polyurethanes, particularly for printing nylon

A(8-E3A3) E(21-B6) F(3-F2, 3-F6A, 3-F8, 3-F10, 73

carpets. (I) are fast to rubbing, light, wet treatments and give even dyeings on stripy nylon. When used with other dyes they are not subject to catalytic fading.

EXAMPLE

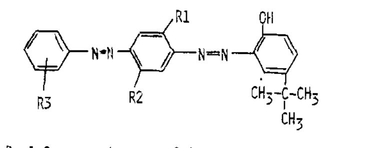
(All pts. by wt.) 27.7 pts. 4-amino-1,1'-azobenzene-4'-sulphonate were dissolved in 200 pts. water at 35-40°C (pH 9-10 with 30% aq. NaOH) then diazotised. The resulting suspension was added over 30 min. to a soln. of 15.2 pts. p-tert butyl phenol in 100 pts. acetone and 200 pts. water keeping pH at 8.5-9 with 10% NaOH. After stirring for 1 hr, the mixt. was acidified and salted out to give the dye

This dyed wool or synthetic polyamide in pure yellow shades with good light- and wet-fastness.(5pp1251)

3-F16A) GB2000171

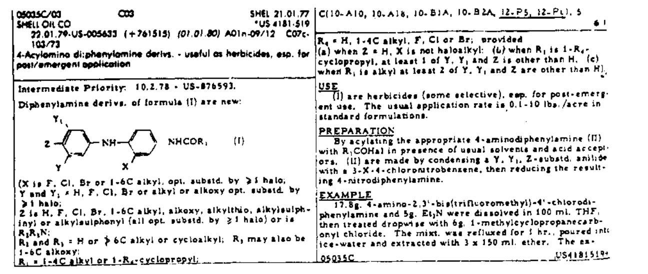
INPUT 2000171

SD



```
R <1-2> = H / ALKYL <1-4> / ALKOXY 1-4 ;
R3 = [<3-4>] SO3 SB ( H / 'EQUIV. OF A NON-
              CHROMOPHORIC CATION' ) .
```

Figure 7.



05035C/03 C03 SHEL 21.01.77
SHELL OIL CO *US 4181-519
22.01.79-US-005633 (+761515) *(01.01.80)* A01n-09/12 C07c-103/73
4-Acylamino di:phenylamine derivs. - useful as herbicides, esp. for post/emergent application

C(10-A10, 10-A18, 10-B1A, 10-B2A, 12-P5, 12-P1). 5 61

Intermediate Priority: 10.2.78 - US-876593.

Diphenylamine derivs. of formula (I) are new:

(I)

(X is F, Cl, Br or 1-6C alkyl, opt. substd. by ≥1 halo;
Y and Y_1 = H, F, Cl, Br or alkyl or alkoxy opt. substd. by ≥1 halo;
Z is H, F, Cl, Br, 1-6C alkyl, alkoxy, alkylthio, alkylsulphinyl or alkylsulphonyl (all opt. substd. by ≥1 halo) or is R_2R_3N;
R_2 and R_3 = H or ≯6C alkyl or cycloalkyl; R_3 may also be 1-6C alkoxy;
R_1 = 1-4C alkyl or 1-R_4-cyclopropyl;
R_4 = H, 1-4C alkyl, F, Cl or Br; provided (a) when Z = H, X is not haloalkyl; (b) when R_1 is 1-R_4-cyclopropyl, at least 1 of Y, Y_1 and Z is other than H, (c) when R_1 is alkyl at least 2 of Y, Y_1 and Z are other than H).

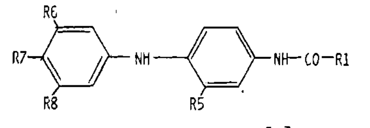
USE
(I) are herbicides (some selective), esp. for post-emergent use. The usual application rate is 0.1-10 lbs./acre in standard formulations.

PREPARATION
By acylating the appropriate 4-aminodiphenylamine (II) with R_1COHal in presence of usual solvents and acid acceptors. (II) are made by condensing a Y, Y_1, Z-substd. anilide with a 3-X-4-chloronitrobenzene, then reducing the resulting 4-nitrodiphenylamine.

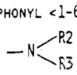
EXAMPLE
17.8g. 4-amino-2,3'-bis(trifluoromethyl)-4'-chlorodiphenylamine and 5g. Et_3N were dissolved in 100 ml. THF, then treated dropwise with 6g. 1-methylcyclopropanecarbonyl chloride. The mixt. was refluxed for 1 hr., poured into ice-water and extracted with 3 x 150 ml. ether. The ex-

05035C US4181519+

```
INPUT 4181519

SD
```

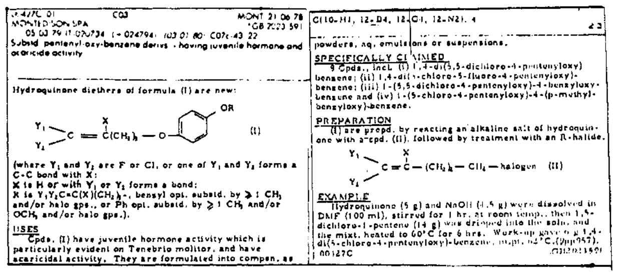


```
R1 = ALKYL <1-4> / CYCLOPROPYL  SB [1] R4  ;
R4 = H / ALKYL <1-4> / F / CL / BR  ;
R5 = F / CL / BR / ALKYL <1-6> SB < 0-> HALO ;
R<6,8> = H / F / CL / BR / ( ALKYL / ALKOXY )
                    SB < 0-> HALO ;
R7 = H / F / CL / BR / ( ALKYL <1-6>/ ALKOXY <1-6> /
      ALKYLTHIO <1-6> / ALKYLSULPHINYL <1-6> /
      ALKYLSULPHONYL <1-6> )  SB <0->  HALO / SD
```



```
                                      ;
R<2-3> = H / ALKYL <-6> /  CYCLOALKYL <-6>;
R3 = ALKOXY <1-6>;
IF R1 = CYCLOPROPYL SB [1] R4  THEN
        RESTRICT <1-> R<6-8> <> H ;
IF R1 = ALKYL THEN RESTRICT <2-> R<6-8>  <> H ;
IF R7 = H THEN RESTRICT R5 <>ALKYL <1-6>  SB <1->  HALO .
```

Figure 8.



04477C/01 C03 MONT 21.06.78
MONTEDISON SPA *GB 2023 591
05.03.79-IT-020734 (+024794) (03.01.80) C07c-43/22
Substd pentenyl-oxy-benzene derivs - having juvenile hormone and acaricide activity

C(10-H1, 12-D4, 12-C1, 12-N2). 4 23

Hydroquinone diethers of formula (I) are new:

(I)

(where Y_1 and Y_2 are F or Cl, or one of Y_1 and Y_2 forms a C-C bond with X:
X is H or with Y_1 or Y_2 forms a bond;
R is $Y_1Y_2C=C(X)(CH_2)_3$-, benzyl opt. substd. by ≥1 CH_3 and/or halo gps., or Ph opt. substd. by ≥1 CH_3 and/or OCH_3 and/or halo gps.).

USES
Cpds. (I) have juvenile hormone activity which is particularly evident on Tenebrio molitor, and have acaricidal activity. They are formulated into compsn. as powders, aq. emulsions or suspensions.

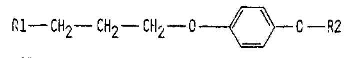
SPECIFICALLY CLAIMED
9 Cpds., incl. (i) 1,4-di(5,5-dichloro-4-pentenyloxy)benzene; (ii) 1,4-di(5-chloro-5-fluoro-4-pentenyloxy)-benzene; (iii) 1-(5,5-dichloro-4-pentenyloxy)-4-benzyloxy-benzene and (iv) 1-(5-chloro-4-pentenyloxy)-4-(p-methyl-benzyloxy)-benzene.

PREPARATION
(I) are prepd. by reacting an alkaline salt of hydroquinone with a cpd. (II), followed by treatment with an R-halide.

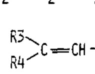
(II)

EXAMPLE
Hydroquinone (5 g) and NaOH (4.5 g) were dissolved in DMF (100 ml), stirred for 1 hr. at room temp., then 1,5-dichloro-1-pentene (14 g) was dripped into the soln. and the mixt. heated to 60°C for 6 hrs. Work-up gave 6 g 1,4-di(5-chloro-4-pentenyloxy)-benzene, m.pt. 64°C.(9pp957).
00127C GB2023591

```
INPUT 2023591
SD
```



```
R1 = SD
```



```
                                  /
     SD
          R3— C≡C—                ;
R<3-4> =  F / CL ;
R2 =  SD
          R1—CH2— CH2—CH2—              /
          BENZYL  OSB <1-> ( METHYL / HALOGEN ) /
          PHENYL  OSB <1-> ( METHYL / METHOXY / HALOGEN ) .
```

Figure 9.

the first *parameter* in the list. For example,

alkyl ⟨1–4⟩

alkyl ⟨3–8⟩ T⟨1–2⟩

Further examples of homologous series terms and *parameter* lists are

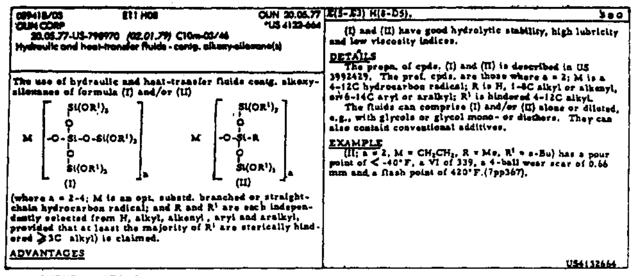
(a) cycloalkyl ⟨10–20⟩ R⟨2-⟩

between 10 and 20 carbon atoms and at least two rings;

(b) alkyl ⟨3–12⟩ 'S' ⟨0–1⟩

between 3 and 12 carbon atoms and 0 or 1 sulfur interruptions;

(c) carbacyclic ⟨6–10⟩ E ⟨1-⟩



08941B/05 E11 H08 OLIN 20.05.77
OLIN CORP *US 4132-664
20.05.77-US-798970 *(02.01.79)* C10m-03/46
Hydraulic and heat-transfer fluids - contg. alkoxy-silanes(s)

E(5-E3) H(8-D5). 3 00

The use of hydraulic and heat-transfer fluids contg. alkoxy-silanes of formula (I) and/or (II)

(I) (II)

(where a = 2-4; M is an opt. substd. branched or straight-chain hydrocarbon radical; and R and R' are each independently selected from H, alkyl, alkenyl, aryl and aralkyl, provided that at least the majority of R' are sterically hindered ≥3C alkyl) is claimed.

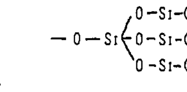
ADVANTAGES
(I) and (II) have good hydrolytic stability, high lubricity and low viscosity indices.

DETAILS
The prepn. of cpds. (I) and (II) is described in US 3992429. The pref. cpds. are those where a = 2; M is a 4-12C hydrocarbon radical; R is H, 1-8C alkyl or alkenyl, or 6-14C aryl or aralkyl; R' is hindered 4-12C alkyl.
The fluids can comprise (I) and/or (II) alone or diluted, e.g., with glycols or glycol mono- or diethers. They can also contain conventional additives.

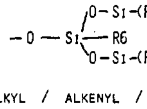
EXAMPLE
(II; a = 2, M = CH_2CH_2, R = Me, R' = s-Bu) has a pour point of < -40°F, a VI of 339, a 4-ball wear scar of 0.66 mm and a flash point of 420°F.(7pp367).

US4132664

```
INPUT 4132664

SD
       R1 —— (R2)M1
R1 = CARBACYCLIC SB  ?  ;
R2 S=  R3 /  R4  ;
R3  =  SD
```



```
                 ;
R4  =  SD
```



```
                 ;
R<5-6> =  H /  ALKYL  /  ALKENYL  /  ARYL  /  ARALKYL  ;
M1  = < 2-4>;
M2  =   3  ;
IF  R2 = R3  THEN  IF  M1 = 2  THEN RESTRICT  <10-> R5  =  ALKYL <3->
                               ELSE IF M1 = 3
                                    THEN RESTRICT <14->R5  =  ALKYL <3->
                                    ELSE RESTRICT <19->R5  =  ALKYL <3->
              ELSE  IF  M1 = 2  THEN RESTRICT <7-> R5  =  ALKYL<3->
                               ELSE IF M1 = 3
                                    THEN RESTRICT <10-> R5  = ALKYL<3->
                                    ELSE RESTRICT <13-> R5  = ALKYL<3-> .
```

Figure 10.

between 6 and 10 carbon atoms and at least one double bond (number of triple bonds not specified). (The term "carbacyclic" is used to indicate acyclic hydrocarbons.)

The assignment statements for R2 and R3 in Figure 5 include homologous series identifiers and *parameter* lists; in the case of R2 the specification of no ternary or quaternary branching atoms is equivalent to the statement in the Derwent Abstract that it is an *n*-alkyl group.

Position Sets. Syntax diagram 12, for *substituent definition element* (of which *substituent value* is a simple case), allows the inclusion of some information about the position(s) of *substituents* being defined. A *position set* at the beginning of a *substituent definition element* indicates the position(s) in the constant structure at which the *substituent*(s) currently being defined may be attached. Thus

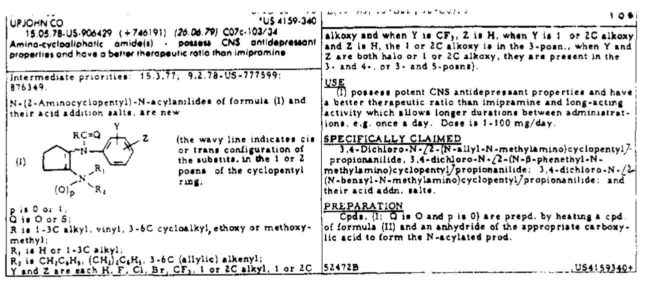
R1 = [2,4] Cl

means that R1 is a Cl group attached in either position 2 or position 4 of the constant structure.

A *position set* following a *substituent* value indicates the position(s) in the *substituent* through which it may be attached to the constant structure. The example in Figure 13 shows that R1 is a nicotinic acid moiety attached through its 2, 4, or 6 position to the 2, 3, 4, 5, or 6 position of the toluene moiety. In both cases the numbering system referred to is the standard one for a nomenclatural term, or whatever numbering of the atoms was employed in the graphic input of the structure diagram in question.

Figures 6 and 7 show patent examples involving *position sets*. The extent to which *position sets* need to be used in GENSAL notations may depend on the facilities available in the graphics system being used for indicating alternative positions of attachment (especially to rings). For example, the Feldmann system being used for our work does not allow the convention of a bond going into the middle of the ring.

Nested Substitution. So far a generic structure has been considered as consisting of a constant part to which variable



UPJOHN CO *US 4159-340
15.05.78-US-906429 (+746191) (26.06.79) C07c-103/34
Amino-cycloaliphatic amide(s) - possess CNS antidepressant properties and have a better therapeutic ratio than imipramine

Intermediate priorities: 15.3.77; 9.2.78-US-777599; 876349.

N-(2-Aminocyclopentyl)-N-acylanilides of formula (I) and their acid addition salts, are new

(I) (the wavy line indicates cis or trans configuration of the substits. in the 1 or 2 posns of the cyclopentyl ring;

p is 0 or 1,
Q is O or S;
R is 1-3C alkyl, vinyl, 3-6C cycloalkyl, ethoxy or methoxymethyl;
R_1 is H or 1-3C alkyl;
R_2 is $CH_2C_6H_5$, $(CH_2)_2C_6H_5$, 3-6C (allylic) alkenyl;
Y and Z are each H, F, Cl, Br, CF_3, 1 or 2C alkyl, 1 or 2C alkoxy and when Y is CF_3, Z is H, when Y is 1 or 2C alkoxy and Z is H, the 1 or 2C alkoxy is in the 3-posn., when Y and Z are both halo or 1 or 2C alkoxy, they are present in the 3- and 4-, or 3- and 5-posns).

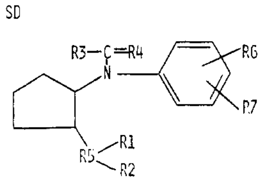
USE
(I) possess potent CNS antidepressant properties and have a better therapeutic ratio than imipramine and long-acting activity which allows longer durations between administrations, e.g. once a day. Dose is 1-100 mg/day.

SPECIFICALLY CLAIMED
3,4-Dichloro-N-/2-(N-allyl-N-methylamino)cyclopentyl/-propionanilide, 3,4-dichloro-N-/2-(N-β-phenethyl-N-methylamino)cyclopentyl/propionanilide; 3,4-dichloro-N-/2-(N-benzyl-N-methylamino)cyclopentyl/propionanilide; and their acid addn. salts.

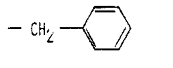
PREPARATION
Cpds. (I; Q is O and p is 0) are prepd. by heating a cpd. of formula (II) and an anhydride of the appropriate carboxylic acid to form the N-acylated prod.

52472B US4159340+

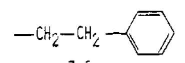
```
INPUT 4159340

SD
```



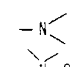
```
R1 = H / ALKYL <1-3>;
R2 = SD
```



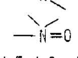
```
                       /
     SD
```



```
                            /
     ALKENYL <3-6>;
R3 = ALKYL <1-3> / VINYL / CYCLOALKYL <3-6> /
         ETHOXY / METHOXYMETHYL ;
R4 = O / S ;
R5 = SD
```



```
                    /
     SD
```



```
                             ;
R<6-7> = H / F / CL / BR / CF3 / ALKYL <1-2> / ALKOXY <1-2> ;
IF R6 = CF3  THEN  R7=H;
IF R6 = ALKOXY <1-2> AND  R7 = H THEN
          RESTRICT R6 = [3] ;
IF 2 R <6-7>  =  HALO / ALKOXY <1-2>
     THEN BEGIN
            RESTRICT  R6 = [3] ;
            RESTRICT  R7 = [4,5]
          END.
```

Figure 11.

PARAMETER IDENTIFIERS :

C	CARBON COUNT
E	ALKENE UNSATURATIONS (DOUBLE BONDS)
Y	ALKYNE UNSATURATIONS (TRIPLE BONDS)
T	TERNARY BRANCH POINTS
Q	QUATERNARY BRANCH POINTS
P	ROOTED PATH LENGTH
R	NUMBER OF RINGS
N	SIZE OF RINGS (NUMBER OF ATOMS)
Z	NUMBER OF HETEROATOMS

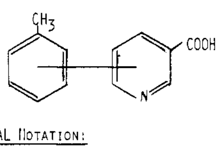
Figure 12.

parts may be attached. In practice, however, these *substituents* may themselves be further substituted by "sub-substituents", varying in nature and position, and this can continue to any level of nesting. GENSAL has facilities, using the mutually recursive syntax diagrams 12 and 13, to show such nested substitution clearly and concisely, the use of parentheses removing any possible ambiguity. The delimiters "SB" and "OSB", respectively, indicate "substituted by" and "optionally substituted by", while "ANDBY" and "ORBY" indicate additional and alternative further substitution. An *integer range* at the start of a *substituent definition element* is a selector and indicates a variable number of occurrences of whatever follows. Thus

R1 = alkyl ⟨4–8⟩ SB (Cl/Br/I) ANDBY ⟨1–2⟩
nitro/(alkoxy ⟨1–6⟩/thioalkyl ⟨1–6⟩) SB
(amino/pyridyl [2] SB [4] (methyl/methoxy))

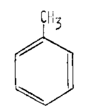
means that R1 can be either (1) an alkyl group of between four and eight carbon atoms, substituted by (a) either Cl, Br, or I and also by (b) either one or two nitro groups or (2) either an alkoxy group or a thioalkyl group (each of between one and six carbon atoms), substituted by either (a) an amino group or alternatively by (b) a pyridyl group attached through its 2 position and substituted in its 4 position by either a methyl or a methoxy group.

"FREE-TEXT" GENERIC STRUCTURE:

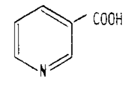


GENSAL NOTATION:

```
INPUT 6789

SD
```



```
                    R1

R1 = [<2-6>]  SD
```



```
      [<2,4-6>] .
```

Figure 13.

The examples in Figures 8 and 9 include *assignment statements* involving bracketing to indicate further substitution, and that in Figure 9 also shows how it may be necessary to alter the way in which the generic structure is expressed in the original specification or abstract in order to encode it in GENSAL.

IF Statements. Many generic structures contain special restrictions on the variety of specific structures that they cover, and GENSAL provides two types of *statement* which allow them to be represented: "IF" *statements* and "RESTRICT" *statements*.

In the IF *statement*, the definitions and RESTRICTions contained in the *statement* following the THEN delimiter are used only in the circumstances when the *condition* is TRUE; the definitions and RESTRICTions in the *statement* following the ELSE delimiter are used only when the *condition* is FALSE. The *statements* in the THEN and ELSE parts may be single or compound (i.e., several *statements* enclosed between BEGIN and END delimiters) and can be *assignment statements*, RESTRICT *statements*, nested IF *statements*, or empty *statements*. The following are simple examples:

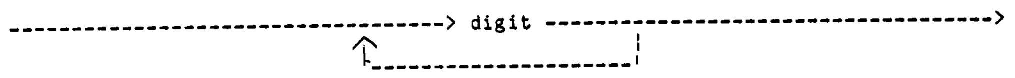
IF R4 = methyl THEN R1 = methyl;
IF R1 = H THEN R2 = H ELSE R2 = Cl/Br/I

Conditions. The syntax diagram for *condition* (No. 16) allows complex *conditions* to be formed, using the Boolean operators AND, OR, and NOT, each component *condition* being a *simple condition*, of the form shown in syntax diagram 15. All *simple conditions* have two sides, separated by a relational operator ("=" or "⟨⟩", meaning "is" or "is not", respectively).

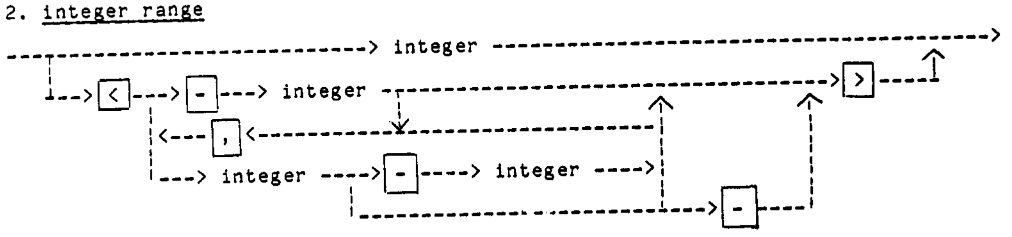
In "definition relations" the right-hand side is a *substituent definition*, of exactly the same form as is used in *assignment statements*, though here it may be abbreviated to a "stand-alone" *position set*, where the chemical nature of the *substituent* is not relevant. The left-hand side consists simply of a *substituent* or *substituent combination* as in

(a) IF R1 = [4] methyl THEN...
(if R1 is a methyl group in the 4 position then...)
(b) IF R2 + R3 = [2/3] THEN...
(if the structure formed by the combination of R2 and R3 is in positions 2 and 3 then...)
(c) IF R4 = alkyl⟨1-6⟩ SB (Cl/Br/I) THEN...

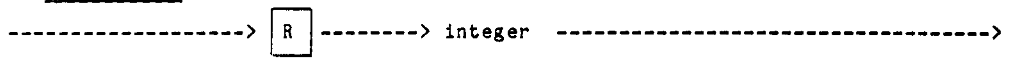
1. integer



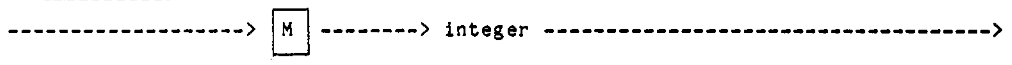
2. integer range



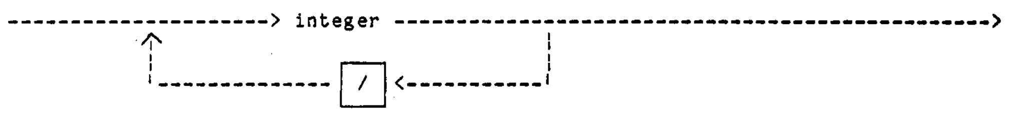
3. substituent



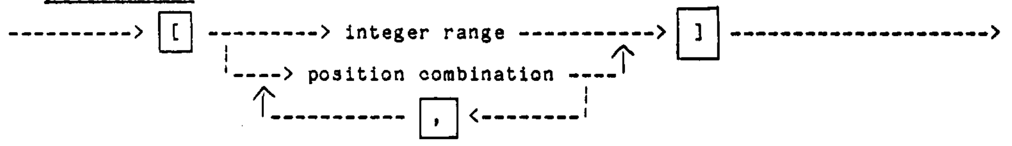
4. multiplier



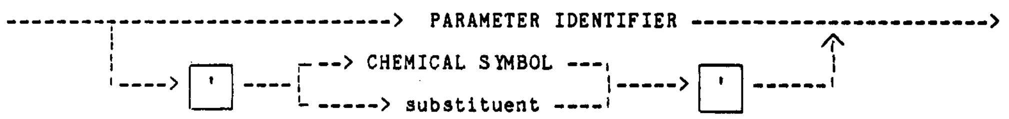
5. position combination



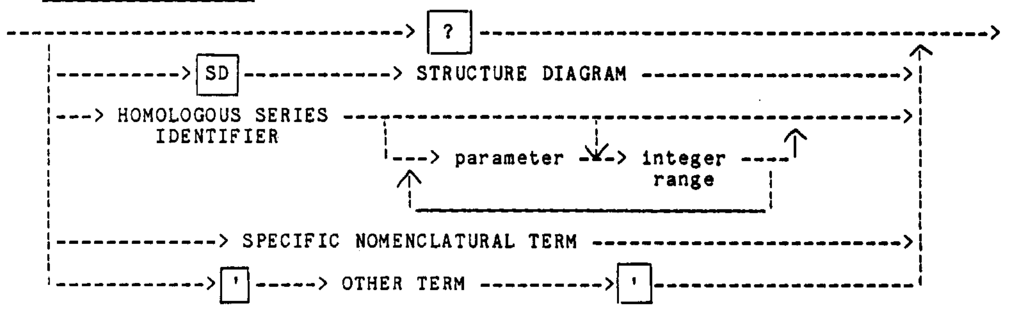
6. position set



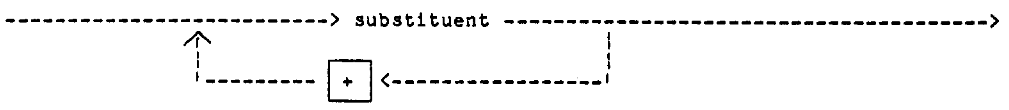
7. parameter



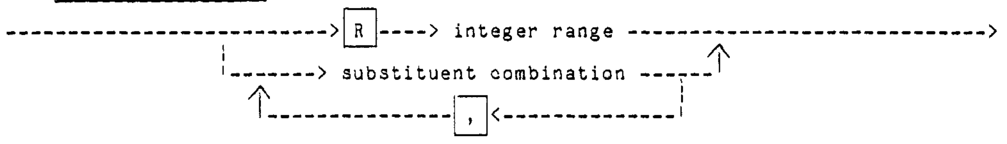
8. substituent value



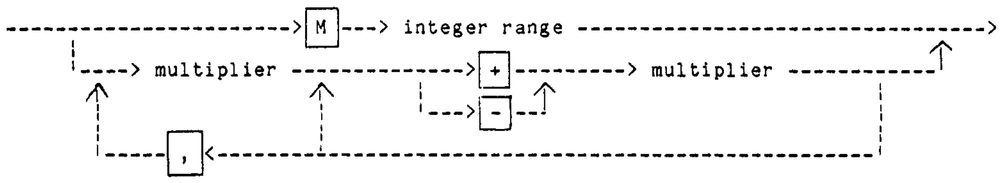
9. substituent combination



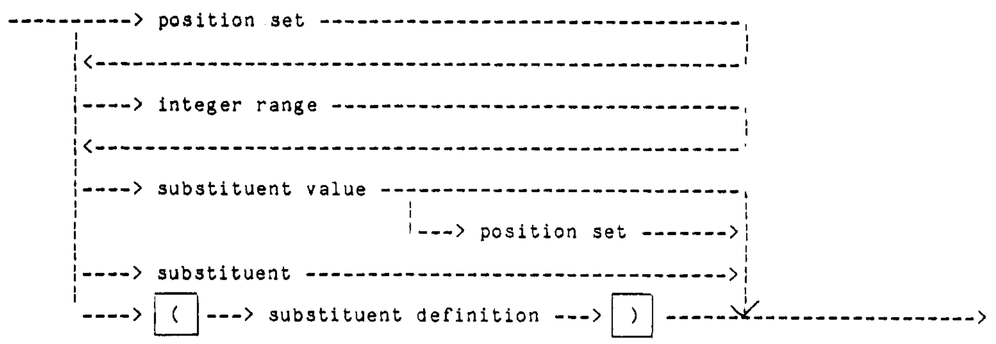
10. substituent group



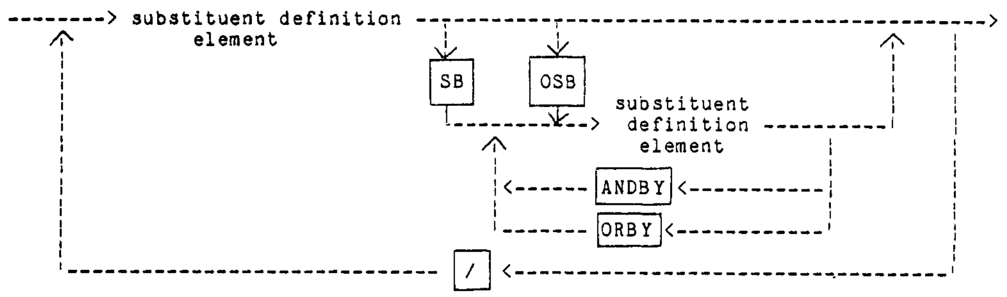
11. multiplier group



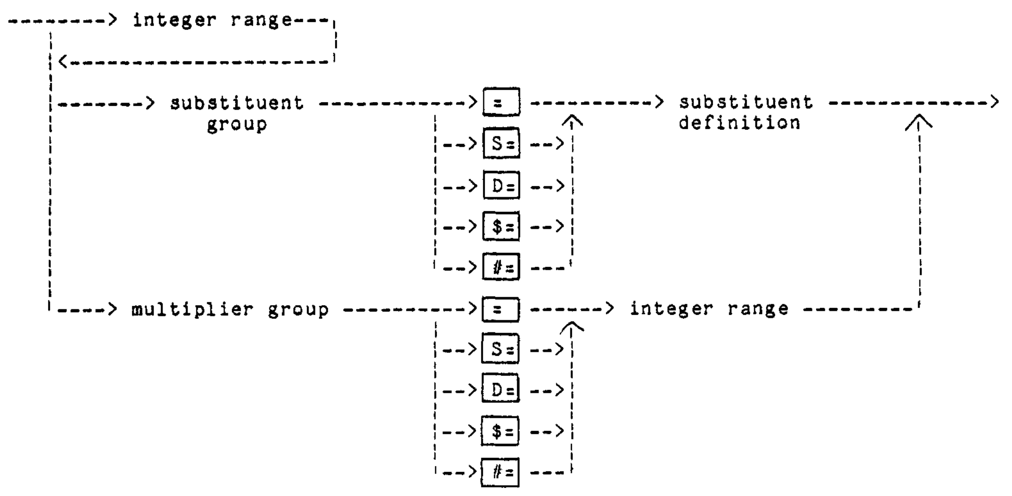
12. substituent definition element



13. substituent definition


substituent definition element
SB
OSB
substituent definition element
ANDBY
ORBY
/

14. assignment statement



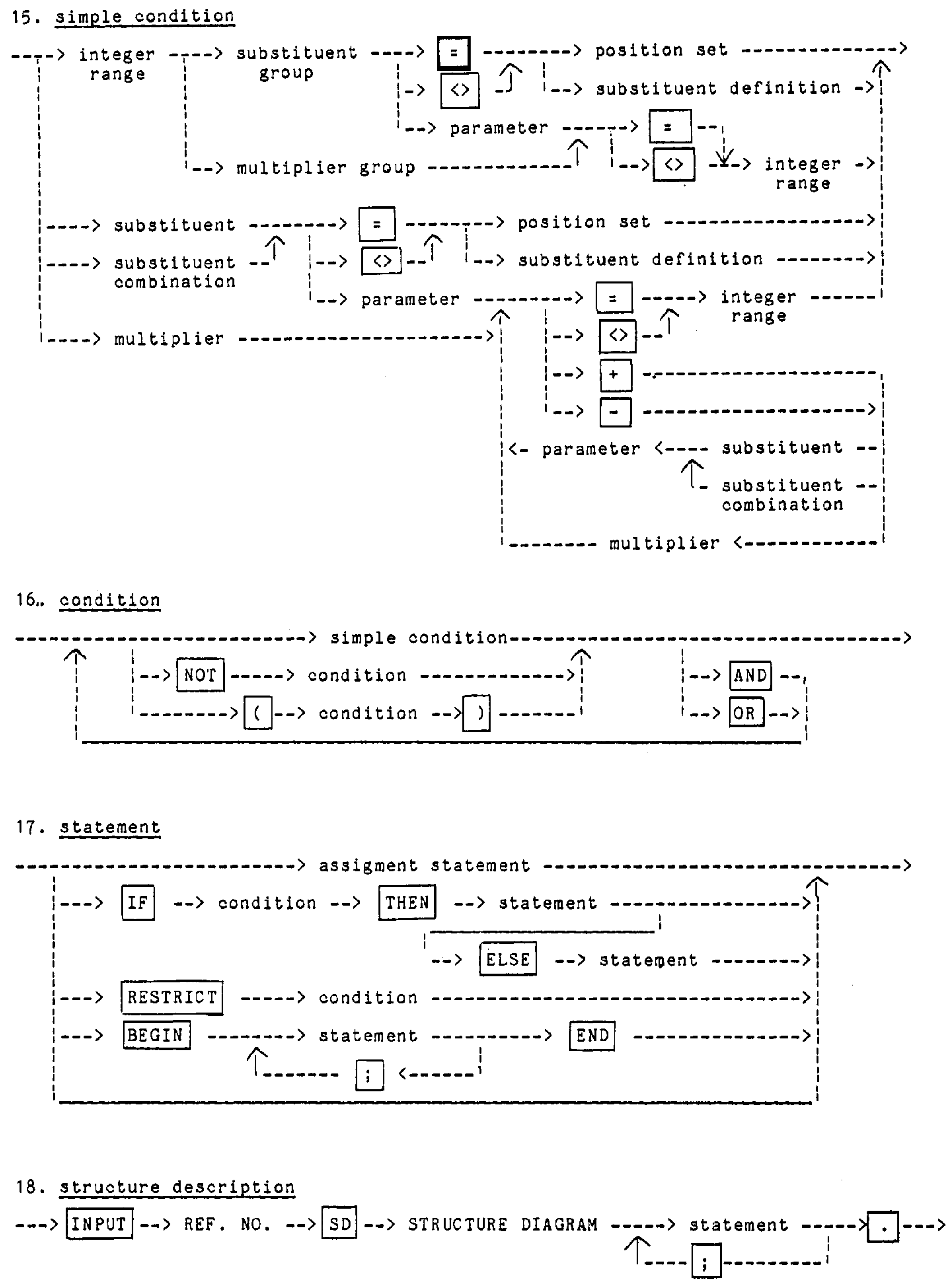

15. simple condition
integer range
substituent group
multiplier group
substituent
substituent combination
multiplier
=
<>
+
-
position set
substituent definition
parameter
integer range
16. condition
simple condition
NOT
condition
(
)
AND
OR
17. statement
assigment statement
IF
THEN
statement
ELSE
RESTRICT
BEGIN
END
;
18. structure description
INPUT
REF. NO.
SD
STRUCTURE DIAGRAM
.

Figure 14.

(if R4 is an alkyl group with carbon atoms between 1 and 6, substituted by either Cl, Br, or I, then. . .).

Integer relations have an *integer range* on the right-hand side of the *condition*, and the left-hand side can consist of various integer terms combined with arithmetic operators, such as *multipliers* and *substituents* with *parameters*:

(d) IF R1 C = ⟨1–2⟩ THEN. . .

(if the carbon count of the homologous series identifier defining R1 is in the range 1–2 then. . .)

(e) IF R2 E = ⟨2-⟩ THEN. . .

(if there are two or more double bonds in the homologous series identifier defining R2 then. . .)

(f) IF M1 = ⟨2–3⟩ THEN. . .

(if M1 is either 2 or 3 then. . .)

(g) IF M1 + M2 + R1 C = 4 THEN. . .

(if the sum of M1 and M2 and the carbon count of R1 is 4 then. . .)

(h) IF R1 C + R2 C = ⟨12-⟩ THEN. . .

(if the sum of the carbon counts of R1 and R2 is greater than or equal to 12 then. . .)

(i) IF R1 + R2 C + M3 = ⟨-6⟩ THEN. . .

(if the carbon count of the combined *substituent* formed by R1 and R2, plus M3 is less than or equal to 6 then. . .).

"Group relations" begin with an *integer range* selector, which operates on the remainder of the left-hand side of the *condition*. If this is a *substituent* group, then the right-hand side will be a *substituent definition* or stand-alone *position set*, as in the definition relations described above:

(j) IF ⟨2-⟩ R⟨1–5⟩ ⟨⟩ H THEN. . .

(if two or more of the *substituents* R1, R2, R3, R4, and R5 are not hydrogen, then. . .).

On the other hand, if the left-hand side is an integer term, such as a *multiplier group* or *substituent group* and *parameter*, then the right-hand side will be an *integer range*:

(k) IF ⟨1–3⟩ M⟨1–5⟩ = 4 THEN. . .

(if 1, 2, or 3 or the *multipliers* M1, M2, M3, M4, and M5 is equal to 4 then. . .)

(l) IF 1 R1 + R2, R3 + R4 C = 3 THEN. . .

(if the carbon count of the group formed by either R1 and R2 or by R3 and R4 (i.e., if the carbon count of one of the two *substituent combinations*) is 3, then. . .).

Only *substituents* and *multipliers* that have already been defined at least once may occur in *conditions*. (GENSAL imposes no restriction on the number of different *assignment statements* each *substituent* or *multiplier* appears in, and all the different definitions given will be regarded as alternative to each other.) However, when an *assignment statement* appears in the THEN or ELSE part of an IF *statement*, the alternatives given in that assignment will be used to the exclusion of all others given outside the IF, when the *condition* is TRUE (for *statements* in the THEN part) or FALSE (for *statements* in the ELSE part).

RESTRICT Statements. In the same way that IF *statements* observe the *statements* in the THEN and ELSE parts of the *statement*, according to whether the *condition* is TRUE or FALSE, so RESTRICT *statements* impose *conditions* on the definitions that have already been made on *substituents* and *multipliers*.

The form of the *condition* is exactly as in the IF *statement*, and thus RESTRICT *statements* appear as in the following examples:

(a) RESTRICT R1 = H

H must have been given as a possible value for R1 in its original definition, and this *statement* eliminates all the other possibilities.

(b) RESTRICT M1 + M2 ⟨⟩ 6

It does not matter what the original definitions of M1 and M2 were; the RESTRICT removes those combinations of possibilities where their sum is 6.

(c) RESTRICT R1 C = ⟨2–3⟩

Here RESTRICT removes those combinations of possibilities where the carbon count is not in the range 2–3.

By suitable use of IF and RESTRICT *statements*, a great many restrictions in generic structures can be very concisely represented in GENSAL, as can be seen from the patent examples shown in Figures 10 and 11.

LIMITATIONS TO THE LANGUAGE

The examples shown in Figures 7, 10, and 11 illustrate some of the current limitations of GENSAL. The Derwent abstract for the generic structure in Figure 7 indicates that certain of the alternatives are "preferred". This is a concept that GENSAL cannot deal with at all; it would, however, be possible to construct a GENSAL notation for this structure in which only the preferred alternatives were shown, and this could be stored alongside the more general notation.

In Figure 10 it is not possible to show the limitations on R5 adequately. The requirement that it be "sterically hindered" cannot be shown at all (unless it be by indicating some branch points in the parameters), and that the majority of the occurrences of R5 must be alkyl⟨3-⟩ can only be shown by exhaustively enumerating all the possible combinations of R2 and M1 in separate IF statements. This is reasonably satisfactory here, but would not be were there a much larger number of possibilities.

Figure 11 illustrates the lack of facilities to show stereochemistry, which is largely a consequence of the absence of stereochemical indicators in the two-dimensional structure representation used by the Feldmann graphics system; it may well prove possible to modify GENSAL to include stereochemical descriptors in *substituent definitions*, treating them in a similar way to *position sets*.

It is not claimed that GENSAL is comprehensive; nevertheless, experience in encoding generic structures from patents suggests that in its present form it is capable of representing adequately the vast majority. However, the possibility is not ruled out that some modifications and extensions may need to be made to certain elements of the syntax (especially for *conditions*) in order to extend the applicability of the language further.

GENSAL SYNTAX DIAGRAMS

In Figure 14 (syntax diagrams), the words and symbols enclosed in boxes are called "delimiters" and are included in GENSAL sentences exactly as they stand. Those in lower-case letters refer to other syntax diagrams, and these terms are shown in *italics* in the text of the paper. Those words in upper-case letters, and not in boxes, are "data items" and are described in the text.

ACKNOWLEDGMENT

Thanks are due to Drs. P. Willett and S. North, members of the AIOPI Patents Study Group, especially Dr. J. Silk, and the staff of Derwent Publications Ltd. for helpful discussions. Thanks are also due to Dr. M. Elder for provision of software and Derwent Publications Ltd. for permission to reproduce abstracts. The support of the Department of Education and

Science through awards of Information Science Research Studentships to J.M.B. and S.M.W. is acknowledged.

REFERENCES AND NOTES

(1) M. F. Lynch, J. M. Barnard, and S. M. Welford, "Computer Storage and Retrieval of Generic Chemical Structures in Patents. 1. Introduction and General Strategy", *J. Chem. Inf. Comput. Sci.*, preceding paper in the issue.

(2) J. E. Hopcroft and J. D. Ullman, "Formal Languages and their Relation to Automata", Addison-Wesley, Reading, MA, 1969.

(3) K. Jensen and N. Wirth, "Pascal User Manual and Report", Springer Verlag, New York, 1975.

(4) N. Wirth, "Algorithms + Data Structures = Programs", Prentice Hall, Englewood Cliffs, NJ, 1976.

(5) E. A. Geivandov, "Language for Notation of Generalized Structures of Organic Compounds Containing Alternative Delocalized Fragments (Markush Structures)", *Nauchno-Tekh. Inf. Ser. 2*, (10), 21-24, 46 (1972).

(6) A. V. Aho and J. D. Ullman, "Principles of Compiler Design", Addison-Wesley, Reading, MA, 1977.

(7) J. Welsh and M. McKeag, "Structured System Programming", Prentice-Hall, Englewood Cliffs, NJ, 1980.

(8) S. M. Welford, M. F. Lynch, and J. M. Barnard, "Computer Storage and Retrieval of Generic Chemical Structures in Patents. 3. Chemical Grammars and Their Role in the Manipulation of Chemical Structures", *J. Chem. Inf. Comput. Sci.*, following paper in this series.

(9) J. M. Barnard, M. F. Lynch, and S. M. Welford, "GENSAL (Generic Structure Language). Instruction Manual", University of Sheffield Postgraduate School of Librarianship and Information Science, in preparation.

(10) R. J. Feldmann, G. W. A. Milne, S. R. Heller, A. Fein, J. A. Miller, and B. Koch, "An Interactive Substructure Search System", *J. Chem. Inf. Comput. Sci.*, **17** 157-163 (1977).

(11) P. A. Machin, J. N. Mills, O. S. Mills, and M. Elder, "CSSR Crystal Structure Search Retrieval", Revised Version, Daresbury Laboratory, Science Research Council, Warrington, England, 1977.

(12) G. W. A. Milne and S. R. Heller, "NIH/EPA Chemical Information System", *J. Chem. Inf. Comput. Sci.*, **20**, 204-211 (1980).

Computer Storage and Retrieval of Generic Chemical Structures in Patents. 3. Chemical Grammars and Their Role in the Manipulation of Chemical Structures

STEPHEN M. WELFORD, MICHAEL F. LYNCH,* and JOHN M. BARNARD

Postgraduate School of Librarianship and Information Science, University of Sheffield, Western Bank, Sheffield S10 2TN, United Kingdom

Received March 25, 1981

A simple topological chemical grammar is developed, and its possible applications to the computer manipulation of chemical structures are discussed. The generative and recognitive capabilities of the grammar are illustrated by examples. The paper concludes by identifying the role of such capabilities in a generic (Markush) structure search system.

INTRODUCTION

A small number of papers has appeared in the literature which report the application of formal linguistic theory to problems encountered in the representation and processing of chemical information. Linguistic considerations were implicit in the early development of systematic chemical nomenclature, and subsequent applications have been made primarily in the continued design of nomenclature and notational systems and their interconversion with other forms of structure representation.[1-7] Garfield[3] used structural linguistics as the basis of an algorithm for generating molecular formulas from a subclass of chemical names. The conversion of systematic names of organic compounds into topological representations has been investigated by Vander Stouw[6] and more recently by Carpenter.[7]

Fehder and Barnett[8] reported early work on the automatic syntax analysis of linear structural formulas. This was proposed both as a preliminary to their conversion into chemical names and as a possible means of substructure search. More recent work along similar lines has been reported by Barker[9] in which a syntax for molecular formulas is described, together with a simple parsing algorithm which enables the correctness of formulas to be verified. It was suggested that such capabilities should find application in interactive computer-based learning and chemical data base query systems.

The role of chemical structure diagrams in the natural language of the chemist is well recognized. Structure diagrams are used both as an aid to written communication and as a means of recording structural information for archival and retrieval purposes. Rankin and Tauber[10] drew attention to the analogies that exist between chemical structure diagrams and formal languages and suggested the value of chemical grammars as a means for structure analysis. In a subsequent paper,[11] two simple structural grammars were presented, a topological and a geometric grammar. As their names suggest, the first treats a chemical structure diagram as a topological graph, ignoring the actual physical dimensions such as bond lengths and angles, while the latter takes account of the geometric arrangement of atoms and bonds in the plane. Underwood and Kanal[12] provided a formalization of the topological grammar described by Rankin and Tauber in terms of a web grammar, a class of graph grammars previously introduced by Pfaltz and Rosenfeld.[13]

A number of applications of topological grammars can be identified. Syntactic analysis allows the "correctness" of structure representations to be verified. Checker programs are used routinely in retrieval systems which employ a notational representation of chemical structure.[14] This capability assumes particular importance in interactive retrieval systems in which the user himself communicates directly with the computer. Online structure registration and query formulation via direct input of structure diagrams are becoming increasingly commonplace with the use of graphics terminals. In the case of compound registration, file corruption can be avoided by trapping invalid structure descriptions as they are entered. Similarly, wastage of machine time resulting from the submission of erroneous query formulations can be eliminated by prior syntax analysis.

A number of other workers have made either implicit or explicit use of linguistic techniques for chemical structure manipulation. The syntax of a language, as will be described, is defined in terms of a number of "productions" or