(20) Olmsted, J. M. "Advanced Calculus"; Appleton: New York, 1956.
(21) Wipke, W. T.; Ouchi, G. I.; Krishnan, S. *Artif. Intelligence* **1978**, *9*, 173-193.
(22) Hobbs, P. D.; Magnus, P. D. *J. Chem. Soc., Chem. Commun.* **1974**, 856-858.
(23) Trost, B. M.; Keeley, D. E. *J. Org. Chem.* **1975**, *40*, 2013.
(24) Wipke, W. T.; Braun, H.; Smith, G.; Choplin, F.; Sieber, W. "Computer-Assisted Organic Synthesis". *ACS Symp. Ser.* **1977**, *61*.

Chemical Inference. 2. Formalization of the Language of Organic Chemistry: Generic Systematic Nomenclature

JOHN E. GORDON[1]

Chemical Abstracts Service, Columbus, Ohio 43210, and Department of Chemistry, Kent State University, Kent, Ohio 44242

Received October 14, 1983

The role and importance of generic nomenclature in the definition and manipulation of structural formula and compound classes in chemistry, chemical information science, and chemical education are discussed. Traditional generic organic nomenclature is surveyed, and a formalization of one traditional language of generic names is presented. Principles of naming generic structural formulas that involve structural variables such as 'R', 'X', etc. are discussed. A formal description of a language for naming and indexing connectivity-variable generic structural formulas of both fixed and variable composition is provided.

MOTIVATIONS TO GENERIC NOMENCLATURE

Indexable Line Notation in 1:1 Correspondence with Generic Structural Formulas. In a recent discussion of languages of generic structural formulas (GSFs), we noted their importance as vehicles for the precise definition and discussion of compound and structural formula (SF) classes.[2] Some of the applications identified were for chemical inference (with and without mechanization), for communication with organic chemistry learners, for discussing substructure search, and for constructing and searching chemical patent claims. Much the same significance attaches to languages of systematic generic names, which stand in the same relation to GSFs as specific systematic names bear to individual structural formulas. And, as with their specific counterparts, generic names (GN) correct one major deficiency of GSFs: GNs are indexable whereas GSFs are not.

Use in Database Searching. Much compound/SF-oriented searching is generic, i.e., directed at compound/SF classes. In addition to much of the patent literature, this includes all searches involving unknown (as yet unisolated) compounds and those involving unknown properties of known compounds. The latter types of searching are carried out as searches for structural analogues of the actual target structures. Chemists routinely employ concepts rich in structural analogies not only in their information-seeking behavior but also in their experimental design and inferential activities. Since structural analogies can be formulated on several different dimensions corresponding to choice of different attributes as analogous, and also in various degrees of strength, a given problem or search may at once involve several different SF classes as analogues for the same unknown compound. Thus, if I wish to find information on the photochemical properties of γ-chloro, α,β-unsaturated amidines, and no such specific information exists, I may wish to search for molecular orbital calculations on unsaturated amidines, for photochemical properties of amidines in general, and so on.

Despite the importance and frequency of such SF class searches, they are easy to carry out only in files possessing either strongly hierarchic organization (e.g., Beilstein) or indexes containing large numbers of SF class enttries.[3] The large bibliographic files most useful for specific SF/compound searching (e.g., *Chemical Abstracts*, *CASearch*) are not so indexed.

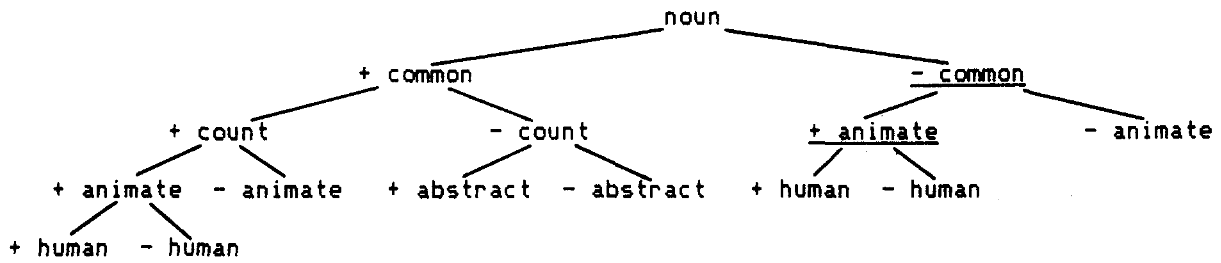
Formalization Exposes Incompleteness, Inconsistency, and Ambiguity. As with all intuitive naming schemes, the traditional generic nomenclature (see below) is difficult to use because it is nonuniform. Thus, while I may call a certain SF class the *aryl alkyl ethers*, others may have discussed or indexed it under *aromatic ethers*, *alkoxy aromatics*, *alkoxy arenes*, *aryloxy compounds*, etc. A second difficulty lies in the existence of a generic nomenclature only for heterocomposite SF classes,[2] not for classes framed in terms of variable connectivity at constant composition. Only in sporadic cases do we have reasonably systematic names for sets of isomeric SFs—even sets of closely related isomers. We generally resort either to ambiguous *specific* systematic names that can be interpreted as naming SF classes, for example, "dichlorobenzene", or to natural language descriptions of the class. Examples are "1-phenyl-3,5-hexanedione and its tautomers", "2-amino-5-sulfo-1-naphthoic acid and its betaines", and "the isomeric decanes".

Sharpening the Chemist's Accuracy of Expression. Generic names share, in a visually less immediate but often more concise form, the role of GSFs as conceptual tools for visualizing, designing, and specifying subclasses of compounds in which some features are constant, others variable. As in most situations involving language use, chemists tolerate considerable levels of ambiguity in the description of SF classes, because in many cases local convention or quick (conscious or unconscious) inference makes the meaning clear. In some situations, however, such ambiguity is not tolerable. These include on the one hand the formulation of index entries and search queries. Equally important, in all formal or informal information transfers that involve learners, ambiguity is destructive of learning because the learner lacks just the chemical intuition that the expert uses to resolve ambiguity.

DESIDERATA FOR SYSTEMATIC GENERIC NAMES

Generic names (1) should have good formal continuity with specific systematic names, (2) should exist hierarchically in 1:1 correspondence with GSFs, (3) should make as explicit as possible (a) the structurally known vs. the structurally

Chart I



unknown parts of a class-defining structure and (b) the secondary structural features present (unsaturation, conjugation, rings), and (4) should support a rational indexing strategy providing reasonably intuitive access to the important SF classes.

CATEGORIAL AND SURFACE STRUCTURE OF CLASS NAMES

A categorization has an underlying logical framework, which we will call the *categorial* structure, and a language of category labels, which we will call the *surface* structure. Terms that denote concepts used in the definition of the categorial structure (e.g., *unsaturated*) often come to expression in the surface language as totally different forms (e.g., 'conjugated' or 'allylic'). Tree structures, such as those in Figure 1a,b, are commonly used to represent categorizations, though not without some ambiguities of interpretation,[5] which often involve interaction between the categorial and surface structures. We speak of the class dominating all the others as the root or the first rank, and we refer to the categories that result from *n* categorization operations on the root as the *n* + 1st rank of the categorization. We speak of the operation of distinguishing daughter categories according to criteria defined by the categorial structure as "interrogating" the parent category. When at every rank the interrogation results in daughter categories that constitute a partition of the parent, the resulting categorization is strictly hierarchic, so that each category is dominated by exactly one predecessor category on the preceding rank.

A tree structure is called a taxonomy if and only if it is strictly hierarchic and if all possible pairs of its classes are either disjoint or mutually related as proper sub- or superset. Care must be taken in constructing surface-language names for classes if such names are to be used in interpreting tree structures. For example, Chart I represents a well-known categorization of English nouns that uses as each of its interrogations a test for the presence of a single feature of the noun. The notation '+ common' is used to indicate the subset (of the predecessor set) of nouns possessing the attribute 'common'. Stewart[5] has claimed that this is not a taxonomy of nouns because the underscored classes, labeled '+ animate' and '- common', are overlapping (not disjoint, not mutual sub- or supersets), on the basis that '+ animate' contains elements both within and outside '- common'. This is fallacious, and it occurs because '+ animate' and '- common' have been taken at face value, whereas they are inadequate descriptive names for the classes, accurate descriptors for which must *conjoin all the relevant features displayed by nodes dominating the node in question.* Thus the underscored classes are actually '- common' and '- common & + animate', from which it is clear that the latter is a proper subset of the former.

There are, however, many ways to express conjunction in a surface language of class names (called *taxons* if the tree is a true taxonomy). The categorizations in Figure 1b use two facets of natural language to accomplish this: (a) cumulation of modifiers, as in 'saturated acyclic', and (b) semantic properties of the lexemes such as those guaranteeing that 'acyclic' is a proper subset of 'aliphatic'. The latter method relies upon unambiguous definitions—an assumption whose shortcomings we will discuss shortly. Except for possible ambiguities of this sort, then, the trees of Figure 1b are taxonomies.

When the subcategorization of a class in a tree structure does not partition that class, so that daughter classes overlap, nodes can be dominated by more than one node on the immediately preceding rank. The categorization is then *crossed.*

When an entity is physically divisible into proper parts, it is frequently named by combining names for the parts. When the part names are built upon their individual categorial structures (which may be quite distinct), the categorial structure of the composite nomenclature will in general include an algebra governing the combinatoric possibilities and any logical dependencies between the parts. The lexicon of the surface langugage of composite names will in general deploy a class of *connectives*, which describe any necessary details of the relationship holding between the parts. The surface syntax of the connectives may be dependent upon syntactic or semantic properties of the name parts that they connect. All of these complications are encountered in the SF domain, since SFs are traditionally analyzed into *functional* and *skeletal* substructures, and the existing, informal language of SF taxons is built around this division. The study of SF categorization will therefore be pursued by categorizing skeletons and functional groups separately and merging the results.

With this background on the form of the categorizations, let us examine the *content* of the SF skeleton categorizations in Figure 1. Beilstein, the great lexicographer of organic chemistry, institutionalized the natural categorization of organic compounds/SFs deemed most useful by turn-of-the-century European chemists; this is shown in Figure 1a. It is a truncated strict hierarchy in which the presence of heteroatoms in acyclic skeletons is not recognized as criterial. [We adopt the conventions of italicizing (in the text) the lexemes of the surface language under consideration (in this case that of SF taxons) and of parenthesizing those nodes in a categorization that are not used in the surface language.]

Figure 1b shows the six rational taxonomies that result from three interrogations of the SF domain with Beilstein's criteria, 'cyclic'/'acyclic' and 'hetero'/'carbo', together with a third common one: 'aromatic'/'aliphatic'.[6] Different orders of application of the same set of interrogations produce categorizations with identical terminal ranks but different branching nodes. Each of these contains (in addition to the root) four branching and five terminal nodes; together they display eight branching and five terminal nodes. All of this (cross-categorized) information can be capture in a single *polyhierarchy* (Figure 1d) an equivalent *pseudohierarchy* (in which all the categories appear as terminal nodes; Figure 1c), or an equivalent generative grammar[7] (Figure 1e).

EXISTING GENERIC NAMES

Traditional Skeletal, Functional Group, and Connective Class Descriptors. Traditional SF class descriptors fall into three

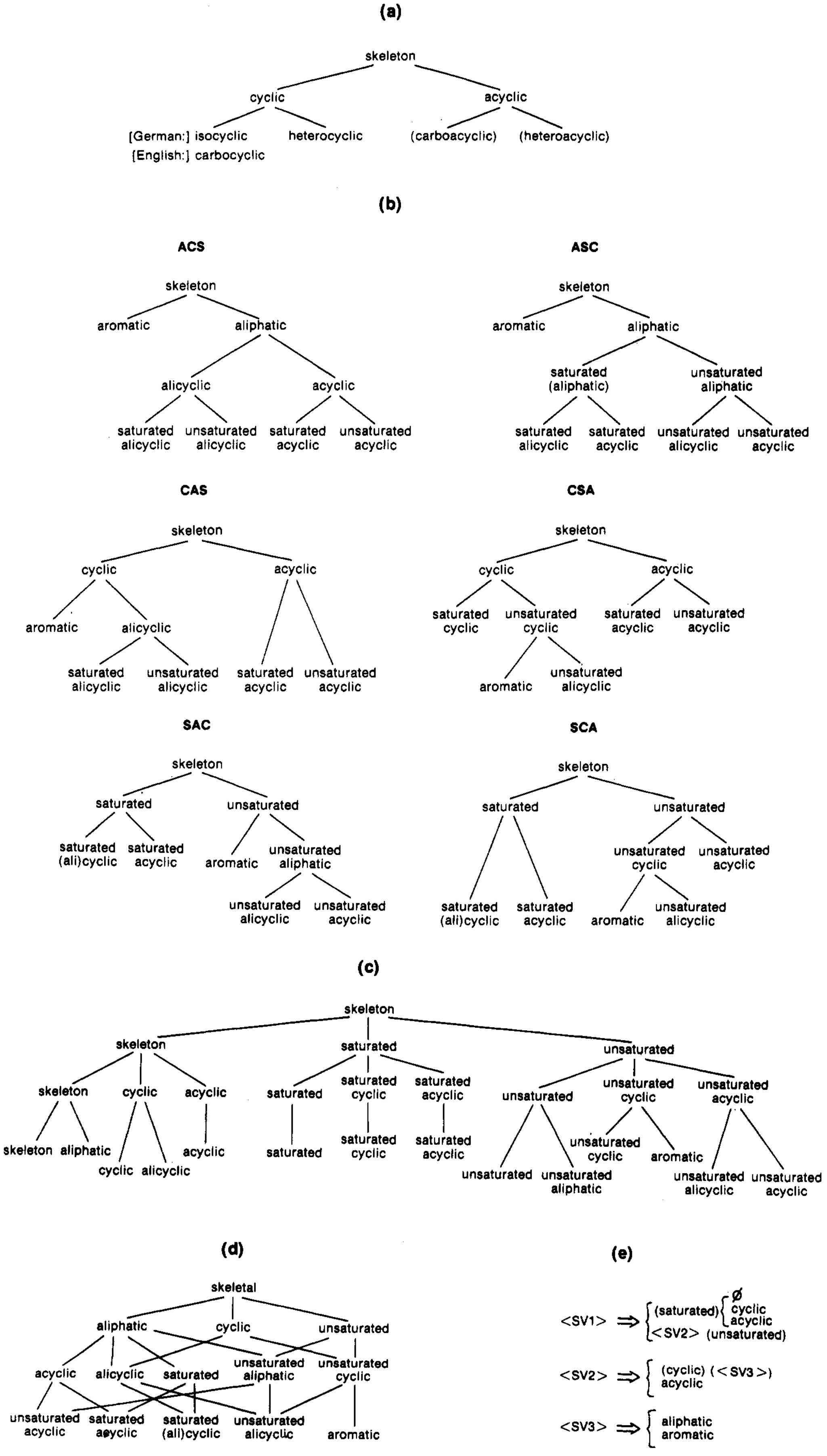


Figure 1. Categorizations of the skeletons of organic structural formulas: (a) Beilstein's categories; (b) hierarchies corresponding to the six permutations of the aromatic/aliphatic (A), cyclic/acyclic (C), and saturated/unsaturated (S) criteria; (c) pseudohierarchic representation of the information in (b); (d) polyhierarchic representation of the information in (b); (e) the information in (b) represented as a generative grammar of SF taxons.

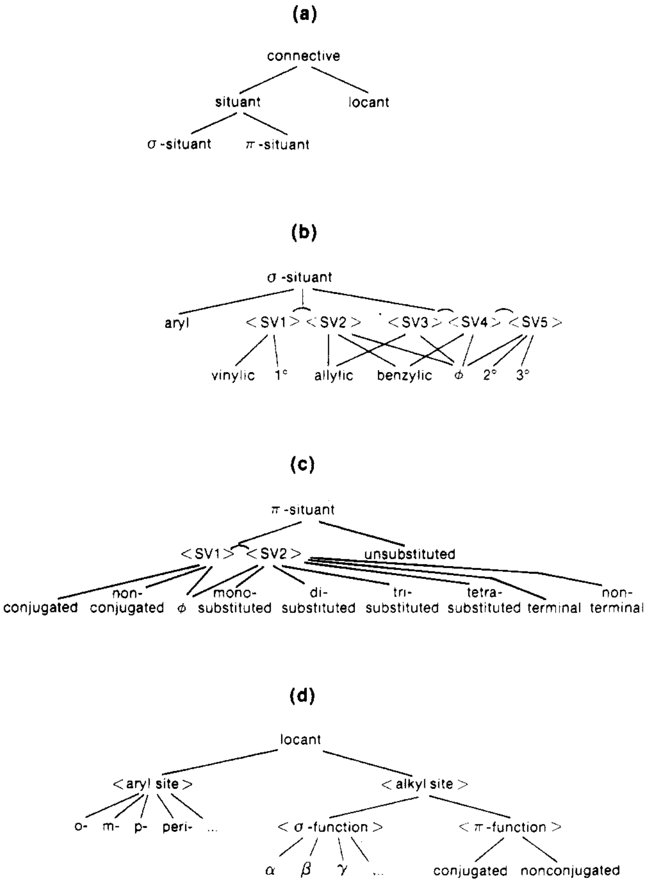


Figure 2. Categorization of the relations between functional groups and skeletons. The hierarchy has been partly dismantled to conserve space.

categories: *skeletal*, *functional group*, and *connective* (describing modes of connection of the other two elements). These terms (a) are distinct from specific systematic nomenclature, (b) are generally of quite long standing and are thus firmly entrenched, and (c) may be combined and modified according to an informal syntax discussed briefly below. This syntax, though incompletely defined and not entirely consistent, supports an informal but rational semantics in which conjunction of descriptors denotes intersection of SF classes. Thus, 'keto' conjoins to produce 'keto alcohol' (or 'hydroxy ketone'), as well as 'aryl, primary alkyl ketone', and these names have reasonably unambiguous denotations (even if the syntax of the latter might be controversial). Operations involving more diffusely defined SF sets are also supported. For example, the implicit class "the SFs of compounds described in the experimental sections of this document" is often ANDed with a traditional functional group descriptor such as 'epoxide' to produce names such as 'the epoxide', unambiguously denoting (in context) the (single) SF in the article with a >C–O–C< substructure.

All of the skeleton descriptors appearing in Figure 1c have been used by chemists, but so have many additional ones. Most of the remainder are captured by including discrimination between the presence/absence of linking (polyvalent) heteroatoms (O, N, S) in the skeleton. This attribute can either replace, say, the *saturated/unsaturated* distinction or go in in addition. The resulting polyhierarchies contain 17 and 83 nontrivial nodes, respectively, many (but not all) of which are observable in chemists' usage.

The traditional functional group descriptors (e.g., *ketone*), in principle, constitute a partition of the SF domain; in fact, there are gaps. Generally speaking, this is a categorization scheme of one rank, motivated by reactivity rather than structural criteria; the functional group taxons distinguish 15–70 characteristic chemical behavior patterns. Some superordinate categories, also functionally motivated (e.g., *carbonyl compounds*, *carboxylic acid derivatives*), are recognized and accurately defined; a few others, e.g., *leaving groups*, are much less precisely characterized.

Chart II

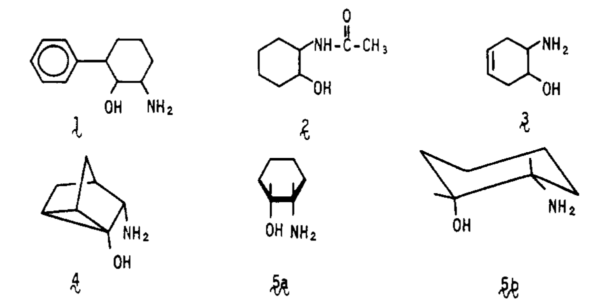


Figure 2 shows an attempt to sort out the connective descriptors in common use. The *generic locants* are quite analogous to the specific locants of systematic (specific) nomenclature; they describe the distance between functional groups, with minimal regard for the structure of the skeleton to which the latter are attached. What we are calling *situants* subcategorize the functional group–skeleton relation; they do so rather finely with respect to the structure of the skeleton.

While all of the descriptors under discussion here are in frequent use, this usage is not well standardized. The main difficulties center on semantic properties of the descriptors: (a) multiple definitions of basic terms (e.g., *aliphatic*) exist, and (b) there is considerable ambiguity about the permissible cooccurrence of situants. We have taken two parallel approaches to working with these traditional generic names. (1) An experimental study of chemists' SF class naming behavior is in progress. (2) In the following section, we describe an arbitrarily formalized syntax for SF class names built on the traditional language considered here. In the remainder of this section, we look at an alternative traditional language of generic SF names.

An Alternative System. One can also form a class name by suffixing 's' to a specific systematic name. A traditional semantics assigns the denotation

(1a) $\{x|x$ is the original SF or an SF derived from the latter by adding substituents$\}$

or perhaps

(1b) $\{x|x$ is any SF containing as a subgraph one of the subgraphs obtainable from the original SF by deletion of nonfunctional hydrogen atom(s)$\}$

Although (1a) and (1b) agree on inclusion of **1** (Chart II) in the class '2-aminocyclohexanols', the concept of a substituent in (1a) is ambiguous under closer examination. Thus, some nomenclaturists would include **2** in the class '2-aminocyclohexanols', whereas many bench chemists would not, on the basis that "substitution" excludes substitutions that turn one functional group into another [e.g., amine → amide; on the other hand, $-CH{=}CH_2 \rightarrow -C(CH_3){=}C(CH_3)_2$ would probably be accepted]. Interpretation (1b) eliminates this problem, but has its own peculiarities, for instance, the inclusion of **3** in '2-aminocyclohexanols', which again violates the bench chemist's reverence for functional–skeletal distinctions. This is an interesting failure of the literal topological model of descriptive chemistry: while >C–C< is indeed a subgraph of >C=C<, interpreted as a multigraph, this equating of the two edges between the 'C's makes no sense under interpretation as a structural formula, in which the σ and π bonds are structurally and functionally distinct. In-

Chart III

```
<name> ::=  ((<SV1>) ((<substituent>)) ((<situant>)) <family>          (1)

<substituent>  ::=  <locant>⁀'-'⁀<group> ((<substituent>))             (2)

                 ⎧ aromatic
<SV1>   ::=     ⎨ ((<SV2>))                                             (3)
                 ⎩ aliphatic ((<SV2>))

                 ⎧ saturated ((<SV3>))
<SV2>   ::=     ⎨ ((<SV3>))                                             (4)
                 ⎩ unsaturated ((<SV3>))

                 ⎧ acyclic ((<SV4>))
<SV3>   ::=     ⎨ cyclic                                                (5)
                 ⎩ ((<SV4>))

                 ⎧ branched
<SV4>   ::=     ⎨                                                       (6)
                 ⎩ straight-chain

                 ⎧ carboxylic acid
                 ⎪ sulfonic acid
                 ⎪ .
                 ⎪ .
                 ⎪ .
                 ⎪ ester
                 ⎪ amine
<family>  ::=   ⎨ alcohol                                               (7)
                 ⎪ thiol
                 ⎪ ether
                 ⎪ halide
                 ⎪ nitro compound
                 ⎪ alkyne/¬ saturated...* & ¬ unsaturated...*
                 ⎪ alkene/¬ saturated...* & ¬ unsaturated...*
                 ⎩ alkane/¬ saturated...* & ¬ unsaturated...*

                 ⎧ sulfo
                 ⎪ .
                 ⎪ .
                 ⎪ formyl
<group>   ::=   ⎨ .                                                     (8)
                 ⎪ .
                 ⎪ hydroxy
                 ⎪ .
                 ⎪ .
                 ⎩ <hal>⁀o

              ⎧ <σ-situant>/(*...alcohol V *...amine V *...halide V
              ⎪      *...ether V *...ester V *...nitro compound V
              ⎪      *...sulfonic acid V *...thiol)
<situant> ::= ⎨                                                          (9)
              ⎪ <π-situant>/(*alkene V *alkyne V *aldehyde V *ketone V
              ⎪      *carboxylic acid V *ester V *amide V *nitrile V...V
              ⎩      *imine)

                ⎧ aryl/[¬ aliphatic...* & ¬ saturated...*]
                ⎪            ⎛⎧ allylic  ⎞
                ⎪ vinylic   ⎜⎨          ⎟/¬ saturated...*
                ⎪            ⎝⎩ benzylic ⎠
<σ-situant> ::= ⎨    ⎛⎧ allylic/¬ saturated...*                        ⎞      (10)
                ⎪ 1° ⎜⎨                                                 ⎟
                ⎪    ⎝⎩ benzylic/[¬ aliphatic...* & ¬ saturated...*]   ⎠
                ⎪ ⎛⎧ 2°⎞
                ⎪ ⎜⎨   ⎟  (allylic/¬ saturated...*)(benzylic/[¬ saturated...* &
                ⎩ ⎝⎩ 3°⎠                ¬ aliphatic...*])

                ⎧ monosubstituted|disubstituted|trisubstituted|
                ⎪      tetrasubstituted/*...alkene
<π-situant>  ::= ⎨ terminal|nonterminal/*...alkyne                       (11)
                ⎪ conjugated|nonconjugated/[unsaturated...* &
                ⎪      (*...carboxylic acid V
                ⎪      *...ester V *...amide V *...aldehyde V
                ⎩      *...ketone V *...nitrile)]

<locant>  ::=  α|β|γ|...|ω/[¬ *...alkane & ¬ *...alkene & ¬ *...alkyne]     (12)

            ⎧ fluor
            ⎪ chlor
<hal>  ::=  ⎨                                                              (13)
            ⎪ brom
            ⎩ iod
```

terpretation (1b) also includes **4** under '2-aminocyclohexanols'; (1a) is again ambiguous.

An alternative, equally consistent semantics gives a considerably more restricted interpretation than (1a) or (1b):

(2) {$x|x$ is an SF obtained by further specification of incompletely defined structural detail in the original SF}

According to interpretation (2), only **5a** or **5b** and the like would be included under '2-aminocyclohexanols'.

FORMALIZATION OF TRADITIONAL GENERIC NOMENCLATURE

While either of the above traditional languages could be standardized and give a formal description, we consider here only the first one, considering it to be the most flexible, the most consistent with the above desiderata, and also nicely complementary to the languages described later in this paper. The rules shown in Chart III sketch the form that a generative grammar for this language might take.

The notation used is Backus–Naur form[8] with two modifications. Alternative replacement values to the immediate right of the '::=' are sometimes separated vertically as

$$\left\{\begin{matrix} a \\ b \\ \vdots \\ c \end{matrix}\right.$$

rather than horizontally (as a|b|...|c). Agreement of name parts is controlled by limting the contexts into which certain formatives may be inserted, as follows: 'v ::= f/e' denotes "syntactic variable v may take the value f if and only if logical expression e takes the value TRUE when evaluated for the string under construction". '*' marks the position of v in the string; '*b' ('b*') indicates a position in the string immediately preceding (following) an occurrence of the formative b; '*...b' ('b...*') indicates a site followed (preceded) somewhere in the string by b. Name parts are separated by spaces unless otherwise indicated by the concatenation sign '⁀'. In those rewriting rules requiring no contextual restriction, absence of '/' abbreviates '/TRUE'. 'SV' abbreviates *structural variable*. The rules are applied in order, such that every structural variable introduced by rules 1–5 and 9 is eventually replaced by formatives via the remaining rules. Recursive rule 2 may be repeatedly applied before passing to rule 3. Parenthesized constituents are optionally retained or deleted on rewriting.

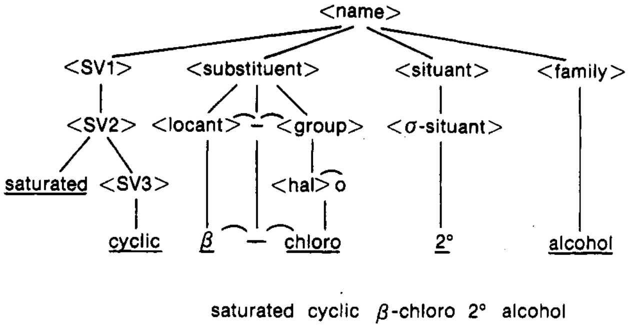


Figure 3. Generation of a name by the "formalized traditional" grammar defined in eq 1–13.

Figure 3 illustrates a sample derivation according to rules 1–13. No claim is made for the completeness of this language (it is not hard to conceive classes of SFs that are not expressible) or for the "naturalness" of its surface forms. Indeed, no real benchmark of traditional usage is available against which the forms generated might be compared. For this reason, refinement of the grammer awaits supplementation of our anecdotal observations with the results of the systematic behavioral linguistic study previously mentioned. This applies both to surface syntactic detail and to choice of lexemes. In the absence of systematically obtained data indicating frequent use of other terms, we have used, in rule 8, substituent names taken from specific systematic nomenclature. These limitations notwithstanding, we believe the names generated by 1–13 to be logically consistent internally and to be rooted in meaningful categorization of the structural elements.

As in natural language,[9] the surface structure may be modified by transformations operating on the products of rules 1–13. One such transformation changes the shape of polysubstituent-containing SF class names as in (14). Here the

aliphatic α-halo α-halo carboxylic acid ==> aliphatic α,α-dihalo carboxylic acid (14)

everyday generic-name language is clearly influenced by systematic specific nomenclature. Other transformations suppress redundancy, as in (15). It is not always clear whether

aromatic nitro benzylic ester ==> nitrobenzylic ester (15)

a feature of the surface language should reside in the phrase-structure rules or in a transformation. Rules 1–13, for example, generate no names for classes of SFs containing more than one instance of the main functional group. This is because we believe that, since traditional naming practice treats the relation between duplicate functional groups the same as the relation between substituent and function (e.g., uses the same locants), the grammar should to the same. Therefore, the duplicate functional group is introduced as a substituent, and the resulting surface form is transformed (16) in close analogy

aromatic sulfo sulfonic acid ==> aromatic disulfonic acid (16)

with the transformation in (14).

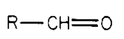
The general sorts of SF classes that are adequately named by the above grammer are those in which the structurally fixed portions are a relatively minor fraction of the whole SF and the overall elemental composition is variable (the SF class membership is *heterocomposite*). There are three sorts of SF classes for which the grammar has poor expressive power: (a) composition-fixed, connectivity-variable (*homocomposite*) SF classes; (b) heterocomposite classes of SFs containing connectivity-fixed skeletal domains; (c) those involving more precise/complex specification of functional group environments than the conjunction of descriptors in the above language can capture.

The remainder of this paper discusses languages tailored to representation of these cases. Since these cases are also the ones that can be represented by generic SFs, these later sections can also attempt to satisfy desideratum 2, above.

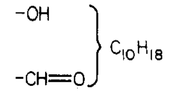
Neither the preceding nor the following material contains any *nomenclature proposals* of the type acted upon by professional societies seeking to maintain stable, normative nomenclatures. It is rather research of the type upon which such proposals are eventually based, and it is submitted for private testing rather than public use. It is confined to SF class names, a topic on which the normative systematic nomenclatures are largely or completely silent.[10] Finally, it represents neither the nomenclatural nor indexing policy of Chemical Abstracts Service or the American Chemical Society, nor nomenclatural principles endorsed by these bodies.

CONNECTED VS. INCOMPLETELY CONNECTED STRUCTURAL FORMULAS

The heterocomposite class consisting of the saturated aliphatic aldehydes is represented by **6** with a structural variable 'R'. We constrain the values that 'R' may assume to structurally completely defined alkyl groups. Hence, **6** is a completely *connected*, though generic, structural formula. On the other hand, **7** is an example of an *incompletely connected*, homocomposite, generic SF—one in which the numbers of 'C', "H', and 'O' signs are fixed, but for only a fraction of these atoms are the connections fully or partly spelled out. Generic SF (GSF) **7** is a well-formed formula of the brace structural formula (BSF) language, which was previously described.[2] While standard nomenclatural principles used in specific systematic names and in the intuitive heterocomposite generic names discussed in the earlier sections of this paper can provide descriptions of the connected parts of SFs like **7**, some new element(s) must be introduced to portray the unconnected portions.



6



7

There are in fact two large and well-known varieties of individual, specific structural formulas that contain points of disconnection: salts and molecular compounds. We propose to adapt aspects of the representation of these substances to the problem of naming BSFs; this is pursued in the final section of this paper. In doing so, we should add one more desideratum to the list given previously: naming of both heterocomposite and homocomposite GSFs should be sufficiently compatible to run under a single, unified algorithm.

A LANGUAGE OF NAMES FOR CONNECTED GSFS CONTAINING NORMALIZED STRUCTURAL VARIABLES

In a fundamental study of languages of generic structural formulas,[2] we (a) reaffirmed the importance of the traditional structural variables 'R', 'Ar', and 'X', (b) provided well-motivated global definitions for them, (c) added a completely general substructure symbol with variable valence ('G'), and (d) provided a formal description of a language of GSFs incorporating these variables. We now explore the rules required to provide generic names for heterocomposite, connected GSFs containing one or more of these variables. Naming of disconnected GSFs containing structural variables is taken up in the section on brace SF naming. While the disconnected generic names considered there bear little *overall* resemblance to systematic specific SF names, some of their syntactic *constituents* are derived via specific systematic nomenclature. More importantly, the connected generic names treated in the present section are designed to be fully compatible syntactic and semantic extensions of systematic specific nomenclature into the generic domain. It is therefore necessary to choose the systematic nomenclatural system with which this compatibility is to be achieved. The decision is made on the basis of need for completely ramified systems of (a) *substructure* naming, treating polyvalent as well as monovalent substructures (groups), and (b) indexing conventions, so that the indexability of the generic names produced may be evaluated. The consequent choice was *9th Collective Index* (9CI) *Chemical Abstracts* (CA) nomenclature.[11]

The structural variables to be treated, and the corresponding surface forms to be admitted to specific substitutive nomenclature, are listed in Table I. As with halogens in specific nomenclature, the variable 'X' never plays the role of principal group and never occurs in names in any but the substituent form, 'halo'. The remaining variables require specialized treatment.

Although 'R' never plays the role of principal group, there are circumstances in which it appears in a name in a form other than 'alkyl': (a) 'R' attached directly to a specific principal group, for example, R–OH. (b) 'R' attached directly

Table I. Formatives in Generic Names of GSFs Containing Structural Variables

			name		
sign	denotation	valence	as principal group	as substituent[a]	as freestanding substructure
R	alkyl (saturated)	1	R-H: 'alkane' (see also Table II)	'alkyl'	'alkyl'[a]
X	halogen	1		'halo'	'halo'
Ar	any aromatic hydrocarbon radical	any	Ar-H: 'arene' (see also Table III)	'aryl'	'arene ⟨VALENCE⟩yl'[b]
G	any substructure	any	'compound'	'substituted'	'group(n)'[c]

[a] Necessarily monovalent. [b] In which ⟨VALENCE⟩ is a syntactic variable taking the values 'di', 'tri', 'tetra', etc. For ⟨VALENCE⟩ = 'mono', the group name 'aryl' replaces 'arenemonoyl'. [c] In which $n \in \mathbf{N}$ is a syntactic variable ranging over the natural numbers. Here, the monovalent case 'group(1)' is not collapsed to the substituent name ('substituted').

Table II. Aliphatic Generic Heading Parents[a]

generic structure	heading parent
R-COOH	alkanoic acid
R-SO_3H	alkanesulfonic acid
·	
·	
R-CO-O-R′	alkyl alkanoate
R-CO-NH_2	alkanamide
·	
·	
R-CO-R(′), R-CO-Ar	alkanone
Ar-CO-Ar	methanone
·	
·	
R-NH_2	alkanamine
R-O-R	1,1′-oxybisalkane
R-O-Ar	see Table III
R-H (-X, -OR′, -NO_2)	alkane
R-Ar	alkyl arene
R-G	alkyl compound
R-Ar-G	alkyl arenecompound

[a] Cyclic analogues follow the following conventions. (1) Cycles within an alkyl group of the parent. The change 'R' => $\overline{CH_2\text{-}(CH_2)_n\text{-}CH}$-, $n \in \mathbf{N}$, corresponds to the root change 'alk' => 'cycloalk' in the heading parent. This nomenclature remains consistent and unambiguous when applied within an alkyl group of structures containing two alkyl groups (illustrations below). The behavior of ketones and imines is regular under this rule provided one considers the skeleton a single alkyl group. Neither this nor traditional generic nomenclature has the means to denote carbocycles remote from the principal group. (2) Cycles spanning two alkyl groups in the parent. The change 'R-G-R(′)' => $\overline{CH_2\text{-}G\text{-}(CH_2)_n\text{-}CH_2}$, which leads to heterocyclic parents taking CA heading parents that are completely diverse with variation in ring size, does not correspond to any global root change in the acyclic parent name. Two sorts of cases arise. (a) The ethers and amines are most easily accommodated with replacement ("a") nomenclature. (b) In the remaining cases, the acyclic parents transform to characteristic traditional names (e.g., 'alkanamide' => 'lactam'). Thus, one has the following: $\overline{CH_2\text{-}(CH_2)_n\text{-}CH}\text{-}SO_3H$, cycloalkanesulfonic acid (case 1); $R\text{-}CO\text{-}O\text{-}\overline{CH\text{-}(CH_2)_n\text{-}CH_2}$, cycloalkyl alkanoate (case 1); $\overline{CH_2\text{-}(CH_2)_n\text{-}CH}\text{-}CO\text{-}O\text{-}R$, alkyl cycloalkanoate (case 1); $\overline{CH_2\text{-}(CH_2)_n\text{-}C}\text{=}O$, cycloalkanone (case 1); $\overline{CH_2\ \text{-}O\text{-}(CH_2)_n}$, oxacycloalkane (case 2a); $\overline{CH_2\text{-}NH\text{-}(CH_2)_n}$, azacycloalkane (case 2a); $\overline{CH_2\text{-}(CH_2)_n\text{-}O\text{-}C}\text{=}O$, lactone (case 2b).

Table III. Aromatic Generic Heading Parents

generic structure	heading parent
Ar-X (-NO_2, -OR, -H)	arene
Ar-OH	arenol
Ar-CH_2OH	arenemethanol
Ar-NH_2	areneamine
Ar-$CH_2CH_2NH_2$	areneethanamine
Ar-COOH	arenecarboxylic acid
Ar-CHO	arenecarboxaldehyde
Ar-CN	arenecarbonitrile
Ar-CH_2COOH	areneacetic acid
Ar-G	arenecompound
Ar-C_x-G	arenemethanecompound, areneethanecompound, arenepropanecompound, etc.

to groups that never play the role of principal group, for example, R–H, R–NO_2, R–F, R–Cl, etc. These occurrences require new, generic heading parent names that are close analogues of the standard parents; they can all be captured by the rule

obtain aliphatic generic heading parents from the specific
9CI name of the structural formula in which R =
$CH_3CH_2CH_2CH_2$– by the substitution 'but' => 'alk'

The resulting parents are shown in Table II. (c) 'R' attached directly to another structural variable, for example, R–Ar and R–G. These require wholly synthetic parent names, which are shown below the dashed line in Table II.

The analogous occurrences of 'Ar' require the generic heading parents shown in Table III. These, and more exotic ones, are generated by the rule

obtain aromatic generic heading parents from the
specific 9CI name of the structural formula in which
Ar = naphthyl by the substitution 'naphthalen' => 'aren'

obtain aromatic generic heading parents from the specific 9CI name of the structural formula in which Ar = naphthyl by the substitution 'naphthalen' => 'aren'

To use the 'R'-, 'X'-, and 'Ar'-containing parents in constructing names requires that the parents be prioritized. The rules of specific nomenclature can be used almost unchanged. 'G' is assigned a class seniority between (aa) "selenium and tellurium compounds" and (bb) "carbon compounds: carbocyclic, acyclic hydrocarbons".[11a]

The new heading parents derived from 'G' are in Tables II and III for those cases in which 'G' occurs in close connection with 'R' or 'Ar'. When 'G' occurs as the sole structural variable, or at a greater remove from 'R' or 'Ar', the combining forms used are (a) '...anecompound', (b) '...enecompound', and (c) '...ynecompound', for 'G' attached to sp^3, sp^2, and sp carbon, respectively. Direct combinations of 'G' with specific functional groups are most simply named by suffixing 'compound' to the function's substituent name: (d) 'carboxy compound' for G–COOH, 'chloro compound' for G–Cl, etc.

The examples shown in Chart IV illustrate whole GSF names involving one structural variable. When 'G' and 'Ar' are directly connected, the new generic parent 'arenecompound' is used—provided it is the senior parent. Thus

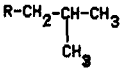
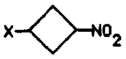
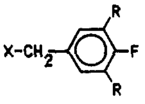
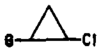
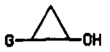
G—Ar—CH_3 methylarenecompound

but

G—Ar—CH=O substitutedarenecarboxaldehyde

When one 'G' and one 'Ar' are connected by a saturated acyclic chain and 'G' is the senior function present, then 'G', 'Ar', and the intervening atoms define a new heading parent according to the 9CI extensions of conjunctive nomenclature.

Chart IV

R-OH	alkanol
	1-alkyl-2-methylpropane
	1-halo-3-nitrocyclobutane
$R\text{-}CH_2\text{-}CBr_2X$	2-alkyl-1,1-dibromo-1-haloethane
	1,3-dialkyl-2-fluoro-5-(halomethyl)benzene
	2-chlorocyclopropanecompound
but	
	2-substitutedcyclopropanol

The form of this name is 'areneALKanecompound', in which 'ALK' takes the values 'meth', 'eth', 'prop', etc. Thus

$G—CH_2—CHX—Ar$ β-haloareneethanecompound

When the last two situations occur simultaneously, as in $G–C_x–Ar–G$, the principle of the larger parent makes the side-chain 'G' senior. Thus

$G—CH_2—CH_2—CH_2—Ar—G$ *ar*-substitutedarenepropanecompound

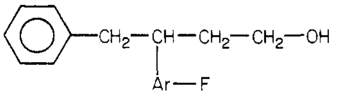
in which '*ar*' is the standard 9CI indefinite aromatic locant.[11b]

The polyvalent 'Ar' and 'G' can bear substituents while playing either their *parent* or their *substituent* role. In the latter case, compound substituents are used in close analogy with specific nomenclature; these are considered below. The former case requires new locants. We adopt the 9CI locant '*ar*' (illustrated in the last example) to indicate substitution on 'Ar', proceeding by analogy with the '*N*-methyl' and '*N*,-*N*-dimethyl' of specific nomenclature, and employ '*g*' to locate substituents on 'G' as in

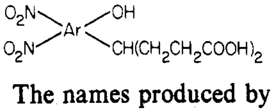
$X—Ar—CH_2—CH_2—CH_2—CH_2—OH$ *ar*-haloarenebutanol

$F—Ar—CH_2—G—OH$ *ar*-fluoro-*g*-hydroxyarenemethanecompound

Substitution on 'Ar' and 'G' *acting as substituents* produces, among others, the complex substituents shown in Chart V. Thus

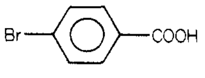


γ-(fluoroaryl)benzenebutanol



4-(hydroxydinitroaryl)heptanedioic acid

The names produced by the above rules appear to be unambiguous and reasonably readable. Use of this language in the mechanized generation of hierarchies of GSFs and generic names in 1:1 correspondence has proved syntactically and semantically satisfactory. Perhaps the weakest aspect lies in names of SFs containing substituents of the 'Br–G–' type. In reading '4-(bromosubstituted)benzoic acid', the mind's eye leaps too readily to



rather than the intended denotation

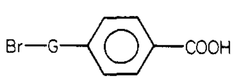
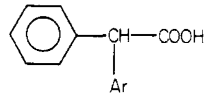


Chart V

X-Ar-	haloaryl
G-Ar-	substitutedaryl
Br-G-	bromosubstituted
R-G-	alkylsubstituted
Ar-G-	arylsubstituted
Br-Ar(G)-	bromosubstitutedaryl
Br-G(Ar)-	arylbromosubstituted
Br-G-Ar-	(bromosubstituted)aryl
Br-Ar-G-	(bromoaryl)substituted

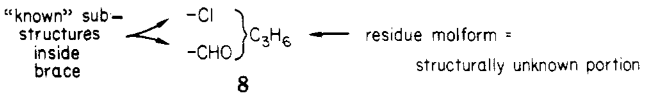
Existing principles fail to provide a choice between the possible heading parents 'benzeneacetic acid' and 'areneacetic acid' for



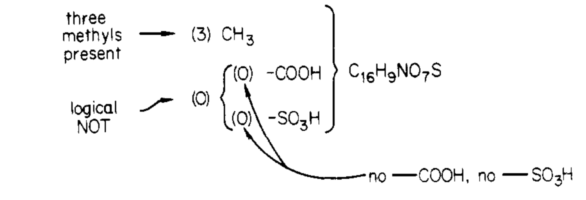
A "principle of the more specific parent" appears to be a well-motivated solution to ambiguities of this type, assigning the name 'α-arylbenzeneacetic acid', rather than 'α-phenylareneacetic acid', to the last structure.

A LANGUAGE OF DISCONNECTED GENERIC NAMES

Isomorphism with Brace Structural Formulas. Brace structural formulas (BSFs), such as **8**, are generic SFs suitable for representing SF classes of fixed composition and variable connectivity. We have demonstrated the completeness and flexibility of expressive power of one BSF language.[2] We now turn to a nomenclature for BSFs. The syntactic schema of the BSF, whose information content the disconnected generic name is to represent, is illustrated in structures **8** and **9**. The



8



9

denotation (extension) of **8** is given by eq 17; the isomer subset

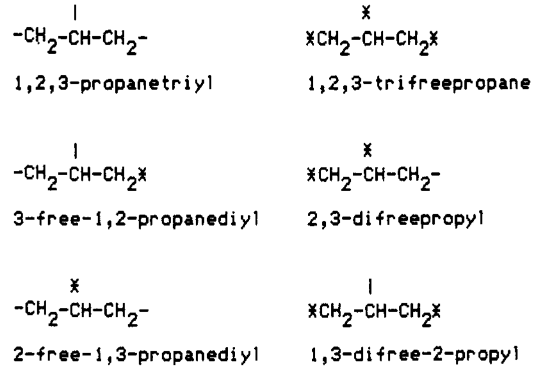
$$K_B(\mathbf{8}) = \{CH_3CH_2CHClCHO,\ CH_3CHClCH_2CHO,\ CH_2ClCH_2CH_2CHO,\ (CH_3)_2CClCHO,\ CH_3CH(CH_2Cl)CHO\} \quad (17)$$

that **9** represents is much larger: all of the isomers of $C_{19}H_{18}NO_7S$ possessing three methyl groups and either a carboxylic acid or a sulfonic acid function.

Open bonds, (–), in the substructures under the brace denote unrestricted variable attachment points. Free valences, '*', denote variable attachments restricted to elements other than hydrogen.

Disconnection in the GSF is to be mirrored by disconnection in the generic name, and we propose to accomplish this by taking into the GN language the dot-disconnection convention

Chart VI



used by *Chemical Abstracts* for denoting disconnection in the structural formulas of molecular compounds and salts.[11d] Appropriate discursive representations of the "known" substructures (SS) within the brace are to be separated by disconnecting dots. These representations are to be as closely related as possible to specific systematic nomenclature. As there appears to be no nomenclature-analogous representation of the residue molform, the simplest course is to take it unchanged into the name and treat it simply as one more block component of the name, dot disconnected from the rest, from which its distinctive format sets it visually apart. The disconnected generic name (DGN) syntax then follows schema 18.

$$\langle\text{SS name}\rangle\text{'.'}\langle\text{SS name}\rangle\text{'.'}...\langle\text{SS name}\rangle\text{'.'}\langle\text{residue molform}\rangle \quad (18)$$

Probably the most fully developed systematic nomenclature for substructures is *Chemical Abstracts* 9CI radical nomenclature, which produces unique names not only for groups but also for substructures with valence (number of open bonds) greater than one. These names serve as a suitable starting point for the body of the DGN.

The components required to complete the definition of the DGN grammar are then (1) rules for modification of 9CI radical names to indicate the partition of indicated substructure valences between open-bond and free-valence types,[2] (2) a convention for representation of quantification under the brace of the BSF, (3) a generative grammar for the DGN language that includes a mechanism for context-sensitized insertion of formatives in order to avert production of ill-formed names whose components are inconsistent, and (4) an algorithm for the unambiguous translation of BSF to DGN. These are considered in turn.

Augmented Radical Names. In 9CI nomenclature, hydrocarbon substructures are named as derivatives of hydrocarbon parent names, via rules that can be paraphrased as use of 'yl' as the quasi-family name for the open bond ('–') "functional group". Normal chain numbering and standard locants for '–' produce names in which the endings 'yl', 'diyl', 'triyl', etc. identify monovalent, divalent, trivalent, etc. substructures. We supplement ths nomenclature with the following rule: Free valences are treated as skeletal substituents on any sort of heading parent, including radicals as parents, with the substituent name 'free'.[12] For example, the substructures derived from propane by removal of one hydrogen from each carbon are shown in Chart VI. We refer to 9CI radical names formed under operation of this additional rule as "augmented radical names".

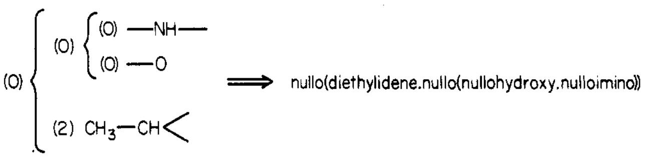
Representation of Quantification, Negation, and Disjunction. In the BSF language constituents of the form '(⟨integer⟩)' occur optionally as prefixes on substructures under the brace. Such a quantifier, prefixed to substructure s_i, denotes the following: if positive or zero, the number of s_is present; if negative, the number of fictitious s_is present due to overlap of substructures; '(0)' prefixed to an inverted (embedded) brace, negation of the (continued) conjunction of (quantified) s_is within the inverted brace. These quantified s_is or s_i conjunctions are rendered in DGN as illustrated in Chart VII. An augmented radical name that bears one of these sorts of prefix is called a quantified augmented radical name. A parenthesized, dot-disconnected string of (quantified) augmented radical names with a prefixed quantifier, such as the last example, is called a logical radical name. Names with the form of logical radical names, but at least one of whose dot-disconnected constituents is itself a logical radical name, are called complex logical radical names, for instance

Chart VII

```
(2) CH3-    ==>   dimethyl

(1) CH3-    ==>   methyl

(0) CH3-    ==>   nullomethyl

(-1) CH3-   ==>   demethyl

(-2) CH3-   ==>   didemethyl
     ┌ (0) -NH-
(0) {              ==>  nullo(nullohydroxy.nulloimino)
     └ (0) -OH
```

Chart VIII

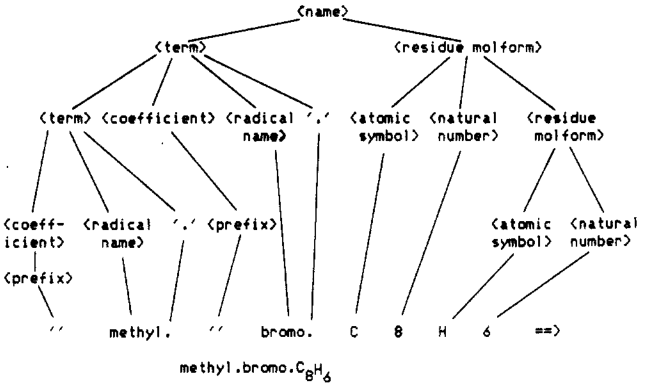
```
<name> ::= <term><residue molform>

           ┌ (<term>)<coefficient><radical name>'.'
<term> ::= {
           └ (<term>)'nullo('(<term>)<coefficient><radical name>')'.'

                      ┌ <atomic symbol><natural number>(<residue molform>)
<residue molform> ::= {
                      └ ϕ

                  ┌ <prefix>('de')
<coefficient> ::= {
                  └ 'nullo'
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
<prefix> ::= '' | 'di' | 'tri' | 'tetra' | ...

<atomic symbol> ::= 'C' | 'H' | 'O' | 'N' | 'F' | ...

<natural number> ::= n ∊ N
                   ┌ x ∊ A/w¥, w ∊ (%, '.', 'nullo', 'di', 'tri',...)
<radical name> ==> {
                   └ x ∊ A/z'de'¥ & ∃y [ŷ ∊ A & d(x)⊂d(y) & #(y)>f(z) &
                            (y...¥ | ¥...y)]
```



Lexicographic ordering of augmented radical name constituents is defined as that specified by paragraph 121 of 9CI,[11c] extended by the convention that all augmented radical names (alphabetized) precede the alphabetically first logical radical name (internally alphabetized).

A Disconnected Generic Name Grammar. In the grammar shown in Chart VIII, syntactic variable names are enclosed in angular brackets, and formatives of the DGN language are in single quotation maarks. The remaining expressions are part of the constraint-to-context apparatus, written in standard mathematical-logical notation, A '{' to the right of the replacement operator '==>' indicates exclusive disjunction—choose one of the expressions within the brace. Parenthesized constituents are optional. Here **A** is the set of augmented radical names, *d*: **A** → **S** is the function that draws substructures given radical names, '%' is a string boundary, '' is the null string, and the function *f* maps values of ⟨prefix⟩ into

Chart IX


⟨name⟩
⟨term⟩
⟨residue molform⟩
⟨term⟩ ⟨coefficient⟩ ⟨radical name⟩ ‘.’
⟨atomic symbol⟩ ⟨natural number⟩ ⟨residue molform⟩
⟨coefficient⟩ ⟨radical name⟩ ‘.’ ⟨prefix⟩
⟨prefix⟩
⟨atomic symbol⟩ ⟨natural number⟩
‘’ methyl. ‘’ bromo. C 8 H 6 ==>
methyl.bromo.C8H6

N as follows: f(‘’) = 1, f(‘di’) = 2, f(‘tri’) = 3, etc. In this notation, ‘⟨constituent⟩ ==> $s/t\ldots{*}\ldots u$’ reads “the syntactic variable named *constituent* may take on value s only when its prospective position in the string under construction lies to the right of an occurrence of ‘t’ and to the left of an occurrence of ‘u’”. The same notation absent the ‘...’ requires t and u as *immediate* neighbors of ‘*’. ‘←’ denotes the relation “is a subgraph of”. ‘$\hat{x}(\ldots x\ldots)$’ denotes the set of individuals x that satisfy the expression ‘...x...’. ‘#(A)’ is the cardinality of set A.

The lexical insertion rules following the dashed line are applied cyclically after cyclic application of the first four (phrase structure) rules has left only occurrences of the syntactic variables ⟨prefix⟩, ⟨radical name⟩, ⟨atomic symbol⟩, and/or ⟨natural number⟩.

A sample DGN derivation is shown in Chart IX. The names produced by this grammar are unambiguous but not unique, because (with the exception of the residue molform) no unique ordering of disconnected name parts has been imposed. This question of uniqueness is taken up in the next section.

Mappings of Brace Structural Formulas onto Disconnected Generic Names. The following algorithm produces DGNs as outputs from BSFs as inputs.

(1) Let the BSF to be named be represented by the schema

$$b = \left.\begin{matrix}(q_1)s_1\\ \vdots \\ (q_i)s_i \\ \vdots \\ (q_n)s_n\end{matrix}\right\} r$$

in which the terms $(q_i)s_i$ may be simple (q_i = integer, s_i = substructure) or complex, in which case $q_i = 0$ and s_i has the form

$$s_i = \left\{\begin{matrix}(q_{i,1})s_{i,1}\\ \vdots \\ (q_{i,j})s_{i,j} \\ \vdots \\ (q_{i,n})s_{i,n}\end{matrix}\right.$$

where the $s_{i,n}$ terms may themselves be complex.

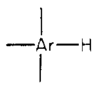
(2) Consider the terms of b. A term situated to the right of an inverted brace, ‘}’, is said to lie within the scope of that brace. An occurrence of ‘}’ within the scope of another ‘}’ is said to be embedded at a level that is one unit deeper than the latter's level of embedding. Terms not in the scope of any ‘}’ are at level 1. (a) If no complex terms are present, go to step 3. Otherwise, begin with the terms under the most deeply embedded instance of ‘}’; let this be called level m. (b) For each term under the currently considered inverted brace, assign an augmented radical name if the term is a simple substructure, s_m, or retrieve the level-m logical radical names produced in a previous cycle through substeps 2a–2e if the term at hand is a complex term. (c) Prefix to each of the names from substep 2b its proper quantifier name = $f(q_{m,i})$, and (d) suffix to each one a ‘.’. (e) Arrange the names from substeps 2a–2c in lexicographic order, concatenate them, delete the final ‘.’, and enclose the resulting string in brackets. The resulting string is a (simple or complex) logical radical name of level $m - 1$. (f) Locate the next unprocessed complex term on level m; if no level-m complex terms remain, decrement m, and locate the first complex term on this next higher level. If necessary, repeat until a complex term is found, in which case go to substep 2a, or, if none is found, go to step 3.

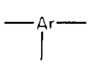
(3) When logical radical names have been assigned to any/all complex terms, those substructures $s_{i,j}$ not in the scope of any inverted brace (i.e., simple terms) are named by (a) assigning an augmented radical name, (b) prefixing the corresponding quantifier name = $f(q_{i,j})$, and (c) suffixing ‘.’.

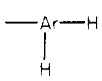
(4) The complete name is assembled from the set of radical names assigned in steps 1–3 to the highest level (simple and/or complex) terms of b, none of which lies within the scope of any internal brace: (a) To each logical or complex logical radical name in the radical name set, concatenate a prefix ‘nullo’ and a suffix ‘.’. (b) Arrange the elements of the radical name set in lexicographic order and concatenate them. (c) Suffix to this string the residue molform r.

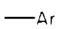
It remains to define the lexicographic order referred to in the algorithm. Absent additional criteria, such as indexability, lexicographic order of radical names in the radical name set can simply be defined in the same way as that used for ordering internal constituents within augmented radical names (above). Thus, we adopt the order prescribed[11c] for substituents in *Chemical Abstracts* index nomenclature and add the convention that all standard (simple, compound, complex) radical names, alphabetized, precede the (internally alphabetized) logical radical names, which are segregated in blocks in order of increasing depth of nesting.

Examples of BSF–DGN pairs produced by the above algorithm and with name parts arranged according to the above ordering rules as shown in Chart X. Several of the examples illustrate the manner in which *heterocomposite* BSFs, which incorporate normalized-variable symbols (‘R’, ‘Ar’, ‘G’), are named. This proceeds by meshing the disconnected generic nomenclature with the augmented traditional generic names discussed in a previous section. Formally, this requires only three new provisions. (1) The BSF in the above algorithm corresponds to the *generalized* definition of a brace structural formula given previously.[2] (2) Tables II and III are annexed to the above definition of radical names. (3) New names are required for intrabrace generic substructures with more than one open bond; these are assembled in the last column of Table I. Two conventions must be provided for their use. (a) They must be modifiable by substitution, as in ‘–G–’ ==> ‘–G–NH_2’ : ‘group(2)’ ==> ‘aminogroup(1)’, which occurs in the third-to-last example. (b) Since **10** is entirely equivalent to **11** and **12** to **13**, etc., open bonds on ‘Ar’ are de facto free valences. We avoid the opacity of using ‘*’ for these valences by adopting the convention of obligatory rewriting all instances of ‘$(–)_m$Ar$(–H)_n$’ as ‘$(–)_{m–n}$Ar’. Note that ‘–G–H’ is not synonymous with ‘–G’, etc.


10

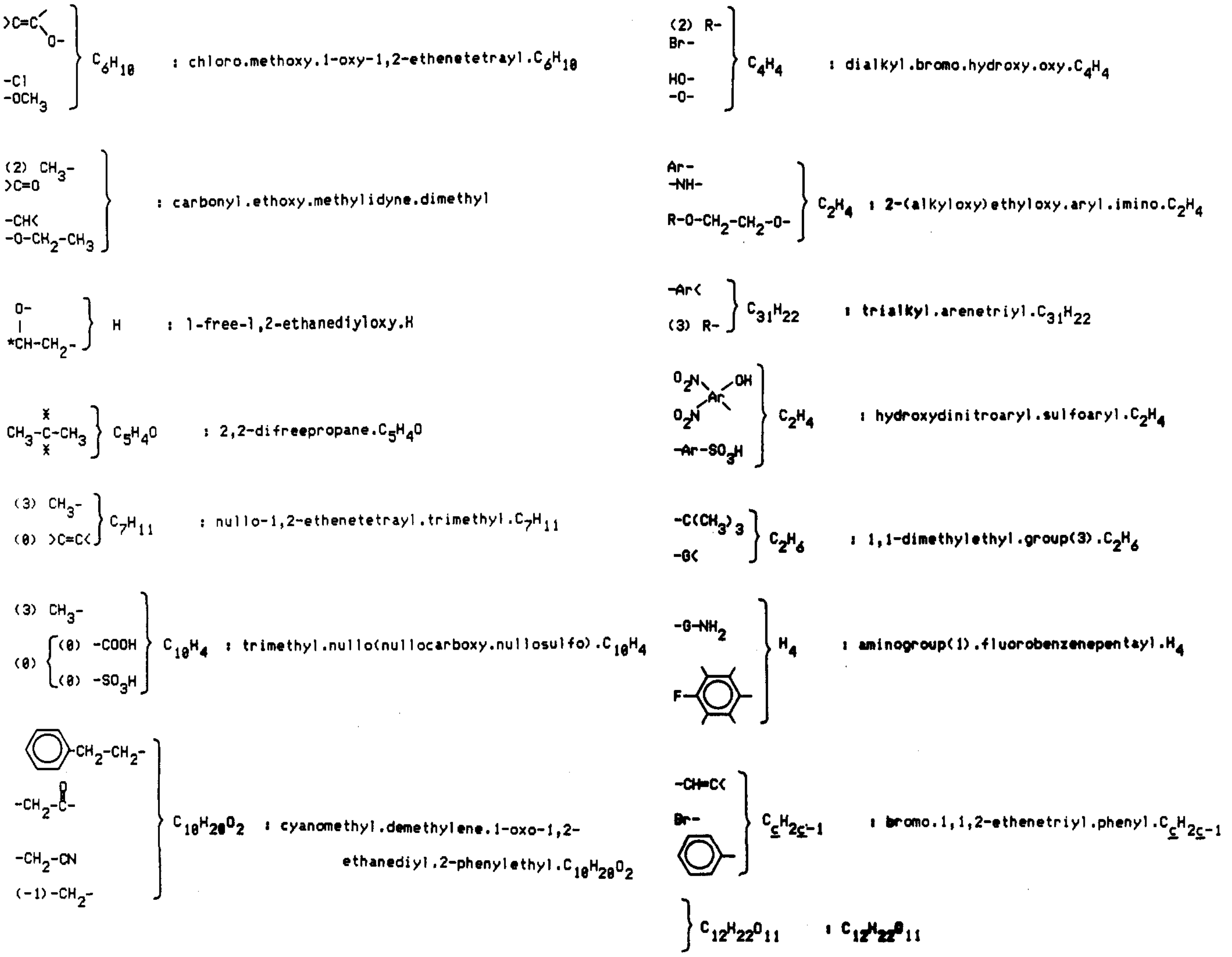

11


12


13

Addressing the question of improving the indexability of these names, we suggest that the following set of criteria for fixing unique, indexable forms of the DGNs is well motivated. The concept, in systematic specific nomenclature, of name

Chart X


C6H10 : chloro.methoxy.1-oxy-1,2-ethenetetrayl.C6H10
: carbonyl.ethoxy.methylidyne.dimethyl
H : 1-free-1,2-ethanediyloxy.H
C5H4O : 2,2-difreepropane.C5H4O
C7H11 : nullo-1,2-ethenetetrayl.trimethyl.C7H11
C10H4 : trimethyl.nullo(nullocarboxy.nullosulfo).C10H4
C10H20O2 : cyanomethyl.demethylene.1-oxo-1,2-ethanediyl.2-phenylethyl.C10H20O2
C4H4 : dialkyl.bromo.hydroxy.oxy.C4H4
C2H4 : 2-(alkyloxy)ethyloxy.aryl.imino.C2H4
C31H22 : trialkyl.arenetriyl.C31H22
C2H4 : hydroxydinitroaryl.sulfoaryl.C2H4
C2H6 : 1,1-dimethylethyl.group(3).C2H6
H4 : aminogroup(1).fluorobenzenepentayl.H4
CcH2c-1 : bromo.1,1,2-ethenetriyl.phenyl.CcH2c-1
C12H22O11 : C12H22O11

inversion to front an index–parent root reasonably gives way, in indexing DGNs, to choice and fronting of a *head radical* in the process of ordering the disconnected name parts. Reasonable criteria for choice of the head radical are based on maximizing its information content. The following rules, applied in order until an umambiguous choice is produced, represent one possibility. Choose the radical that (a) contains the largest number of non-'H' atomic symbol tokens, (b) has the largest number of open bonds plus free valences, (c) has the largest number of acyclic heteroatoms, (d) has the largest number of most preferred acyclic heteroatoms, (e) has the largest number of multiple bonds, (f) has the largest number of free valences, and (f) comes earlist in lexicographic order.

We note in passing that the DGNs are ideally suited for use in a permuted term index in which permutation would range over all radical name parts of the DGN but presumably not the residue molform portion, which contains no structural information and contains compositional information only by difference.

ACKNOWLEDGMENT

This material is based upon work supported by the National Science Foundation under Grant SED 8020159, by Chemical Abstracts Service, a division of the American Chemical Society, and by Kent State University. I thank Joyce Brockwell for numerous suggestions and Chemical Abstracts Service for the opportunity to spend the 1980–1981 academic year as a CAS Visiting Research Scientist. Suggestions of an expert and painstaking referee improved the accuracy and consistency of the article substantially.

REFERENCES AND NOTES

(1) Chemical Abstracts Service Visiting Research Scientist, 1980–1981; correspondence should be directed to the Kent State University address.

(2) Gordon, J. E.; Brockwell, J. C. "Chemical Inference. 1. Formalization of the Language of Organic Chemistry: Generic Structural Formulas". *J. Chem. Inf. Comput. Sci.* **1982**, *23*, 117–134.

(3) Beilstein facilitates (manual) SF-class scanning for analogies on the basis of homology and functional substitution but not on variable unsaturation. For further commentary on SF-class search difficulties, see references 2 and 4.

(4) Lynch, M. F.; Barnard, J. M.; Welford, S. M. "Computer Storage and Retrieval of Generic Chemical Structures in Patents. 1. Introduction and General Strategy". *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 148–150.

(5) Stewart, A. H. "Graphical Representation of Models in Linguistic Theory"; Indiana University Press: Bloomington, IN, 1976; Chapter 1.

(6) Definitions of the criteria are as follows: 'cyclic' = containing at least one ring; 'acyclic' = containing no rings; 'hetero' = containing at least one polyvalent atom other than C; 'carbo' = containing no polyvalent atoms other than C; 'aromatic' = containing at least one aromatic ring; 'aliphatic' = containing no aromatic rings.

(7) This grammar does not define the surface syntax of the object language of SF taxons (which is taken up in the following section); it is a representation of the underlying structure of the categorization.

(8) Ralston, A., Ed. "Encyclopedia of Computer Science"; Van Nostrand-Reinhold: New York, 1976.

(9) Akmajian, A.; Heny, F. "An Introduction to the Principles of Transformational Syntax"; MIT Press: Cambridge, MA, 1975.

(10) A refereee has pointed out that a working party of the IUPAC Nomenclature Commission on Organic Chemistry is studying the nomenclature of class names.

(11) "Index Guide"; Chemical Abstracts Service: Columbus, OH, 1977; Appendix IV, (a) paragraph 106, (b) 120, (c) 121, and (d) 198.

(12) There is a proposal in the literature [Ericson, L. G.; Cutten, D. R. "A Simple Nomenclature for Complex Organic Free Radicals". *Bull. Chem. Soc. Jpn.* **1967**, *40*, 2974-2975] to denote an unpaired electron (a free radical site) by the prefix "keno". Since there is a chemically substantive distinction between a free radical site and a free valence, we have denoted the latter by the improvised "free" rather than "keno".

ACS Committee on Nomenclature: Annual Report for 1983

KURT L. LOENING

Chemical Abstracts Service, Columbus, Ohio 43210

Received March 1, 1984

Nomenclature committees, both national and international, were very active in 1983, resulting in substantial progress in many different fields. A summary of the more important meetings and accomplishments follows.

The *ACS Committee on Nomenclature* held its annual meeting at CAS in November.† Progress of the work of the divisional committees and international commissions was reviewed. The committee recommended a format for the periodic table of the elements that will avoid the existing confusion in designation of the subgroups. The recommended format is the result of 3 years of study and solicited input, such as at the Symposium on the Periodic Table at the Seattle meeting cosponsored by the Committee. The format will be announced in the 1984 February issue of the *Journal of Chemical Education*. Long-range plans of the Committee were submitted and approved by the Long-Range Subcommittee of the Council Policy Committee. Regrettably, the official appointment in 1982 of all editors of ACS journals as ex officio members of the Committee was revoked in 1983, thus returning communications with the editors to a more informal basis. To establish closer contact with interested individual ACS members, the Committee held two successful open meetings in Seattle and Washington and plans to hold two more at the national meetings in 1984. As a result of the Committee's effort to involve more divisions in its work, the Divisions of Nuclear Chemistry and Technology and Chemical Information were represented at the annual meeting for the first time. Presentations stressing the importance of nomenclature and the role of divisional committees in it were given to the Executive Committees of the Divisions of Industrial and Engineering Chemistry and Environmental Chemistry. Closer liaison with nomenclature groups in disciplines related to chemistry continues to be pursued; for example, the Committee is now represented on the Nomenclature Committee of the Council of Biology Editors. Efforts to contact appropriate groups in physics, geology, and mass spectrometry are in progress. The promotion of and input into International Union of Pure and Applied Chemistry (IUPAC) recommendations is, as always, a primary objective of the Committee. As part of the Committee's efforts in the area of public relations, an international nomenclature symposium is being organized for the Honolulu meeting in 1984. The Subcommittee on Chemical Pronunciation continues to be active.

The *IUPAC Interdivisional Committee on Nomenclature and Symbols* (IDCNS) continued to function effectively this year. It held its annual meeting in Lyngby, Denmark, in August. In addition to the IUPAC publications listed in the appendix, specific documents in process and thus not yet recorded in this appendix deal with the following topics: flame emission and absorption spectroscopy, amino acids and peptides, in situ microanalysis, molecular luminescence spectroscopy, etc. A complete list of IUPAC-approved glossaries has been compiled. The key role of IDCNS in the revised IUPAC publication procedure has been codified.

The *IUPAC Inorganic Nomenclature Commission* met in August in Lyngby. Topics included neutral molecules and compounds, ions and radicals, rings and chains, polyhedral clusters, solid-state chemistry, isopoly- and heteropolyanions, oxo acids, inorganic polymers, and stereochemical nomenclature. These topics were discussed in the context of providing a revision of the 1970 edition of the *Red Book*.

The *IUPAC Organic Nomenclature Commission* met in August in Lyngby. The Commission continued its study of the reorganization and revision of the present rules according to a more logical arrangement (Section R) and of a more systematic long-range approach (Section G). In connection with Section G, several specific projects are under way: nodal nomenclature, radial nomenclature, "inorganic" ring nomenclature, nomenclature for delocalized ions and radicals, and nomenclature of oxo acids. The λ-convention published provisionally in January 1982 was approved as final recommendations, 1983. The fully approved revision of the Hantzsch-Widman system of nomenclature for heteromonocycles was published in February 1983. A document extending the generation of names for numerical prefixes beyond 200 was published as provisional rules in August 1983. Initial drafts of documents on a convention for describing rings and ring systems with cumulative double bonds and hydride names for nonmetal hydrides were reviewed. Documents on classical ions and radicals, cyclophanes, and nodal numbering are well advanced. Subjects under study related to the Section R effort include revision of the Section E rules (Stereochemistry), Section F (Natural Products), recommendations on indicated hydrogen, numbering priorities for unsaturation and hydro prefixes, documentation of principles of fusion nomenclature, and names for acid suffixes and parent compounds.

The *IUPAC Macromolecular Nomenclature Commission* also met in August in Lyngby. The commission completed its work on the nomenclature and symbolism of copolymers and on the classification and family names of polymers; these documents are expected to appear in print in 1984. The Commission is continuing its work on (a) subsidiary definitions of terms relating to polymers, (b) definitions for physical properties of polymers, (c) substitutive nomenclature for reacted polymers, and (d) interpenetrating polymer networks. The commission is planning to combine all its recommendations into book form.

In *biochemical nomenclature* both JCBN and NC-IUB met jointly in Tegernsee in May. A major effort was directed toward preparation of a new larger edition of the IUB book

† Abbreviations used, not identified in the text, are ACS, American Chemical Society; CAS, Chemical Abstracts Service; JCBN, Joint Commission on Biochemical Nomenclature; and NC-IUB, Nomenclature Committee of International Union of Biochemistry.

0095-2338/84/1624-0092$01.50/0