

- Stereochemical Notation Abilities," *J. Chem. Doc.*, **10**, 75-81 (1970).
- (3) Bowman, C. M., Landee, F. A., and Reslock, M. H., "A Chemically Oriented Information Storage and Retrieval System. I. Storage and Verification of Structural Information," *J. Chem. Doc.*, **7**, 43-47 (1967).
- (4) Sorter, P. F., Granito, C. E., Gilmer, J. C., Gelberg, A., and Metcalf, E. A., "Rapid Structure Searches Via Permuted Chemical Line-Notations," *J. Chem. Doc.*, **4**, 56-60 (1964).
- (5) Granito, C. E., Becker, G. T., Roberts, S., Wiswesser, W. J., and Windlin, K. J., "Computer-Generated Substructure Codes (Bit Screens)," *J. Chem. Doc.*, **11**, 106-110 (1971).

Automated Conversion of Chemical Substance Names to Atom-Bond Connection Tables

G. G. VANDER STOUW,* P. M. ELLIOTT, and A. C. ISENBERG

Chemical Abstracts Service, The Ohio State University, Columbus, Ohio 43210

Received August 1, 1974

Chemical Abstracts Service (CAS) has developed a computer program for converting systematic names of organic compounds into atom-bond connection tables of the type input to the CAS Chemical Registry System. This program, called the nomenclature translation program, is designed to process names which are based on the word roots and punctuation conventions used in CA Index nomenclature. Both inverted and uninverted names can be processed if they are specific and unambiguous. Inconsistent or ambiguous names will be rejected, with appropriate diagnostics, as will names containing features not recognized by the program. The translation program is currently being installed as part of a comprehensive name editing system.

Chemical substances may be described in several ways, including both structural diagrams and a variety of chemical names. The names used for a given substance may include both "systematic" names, which are constructed from commonly understood nomenclature fragments that correspond to fragments of the structural diagram, and also other names which do not describe the structure of the substance to which they refer. For example, the substance described by the simple structural diagram $\text{N}\equiv\text{C}-\text{CH}_2-\text{CH}_2-\text{C}\equiv\text{N}$ has been referred to not only by structurally descriptive names including "butanedinitrile", "succinonitrile", and "1,2-dicyanoethane", but also by trade names such as "Dinile", "Deprelin", and "Suxil".

The Chemical Substance Index to *Chemical Abstracts* (CA) brings together, under a single name, all CA references to a particular chemical substance which has been selected as a CA Index entry regardless of the various names used in the original documents. The substance appears in the Index at the CA Index Name, a "canonical" name derived by the application of a rigorous and comprehensive set of systematic name selection rules. Preparation of CA Indexes is supported by the Chemical Registry System, a computer-based system which links the structure and various names of a substance; this system included approximately 2.7 million substances at the beginning of 1974. The address of a substance in this System is the CAS Registry Number, a unique identifying number which is associated with a canonically numbered atom-bond connection table in the system's structure file^{1,2} and with the CA Index Name and other names for that substance in the Registry nomenclature file.³

There are two basic routes for retrieval of substance information from the CAS Registry files (see Figure 1). One is "name match", in which a name is compared against the contents of the name file; if a "match" occurs, the Registry Number is retrieved. The other basic route is structure registration, in which a keyboarded structural diagram is con-

verted to the canonical connection table which is matched against the structure file to retrieve the Registry Number. If no "match" occurs during structure registration, *i.e.*, the substance is new to the file, a new Registry Number is automatically assigned. The retrieved Registry Numbers can then be used to retrieve the CA Index Name and any other names from the Registry name file.

Although the Registry System links the names and structural representations of a substance, name and structure have not previously been directly interconvertible. We report here the development of a computer program for converting chemical names into connection tables, a process we call "nomenclature translation".[†] As illustrated by the dotted lines in Figure 1, this process provides an alternate method of structure registration by allowing a new substance to be input *via* a structurally descriptive systematic name instead of only as a connection table taken from a structural diagram.

There are two major potential applications for the use of nomenclature translation as an entry to Registry processing. One is for entering new substances for which systematic names are available into the CAS Registry structure file, bypassing the need for input of structural diagrams. The other application is to verify that the structural records and the CA Index Name on file for a given substance are fully consistent. Processing the CA Index Name for a substance by nomenclature translation, followed by registration of the resulting connection table, should lead to retrieval of the Registry Number previously assigned to that substance. If the expected number is not retrieved, an inconsistency between the name and structure records on file is indicated and the records in error must be identified and corrected. Nomenclature translation thus can provide a powerful tool for use in a system for editing CA Index

[†] The conversion of connection tables to names, called "nomenclature generation", has been described for bridged ring systems^{4,5} and is the subject of continuing CAS investigation.

* To whom correspondence should be addressed.

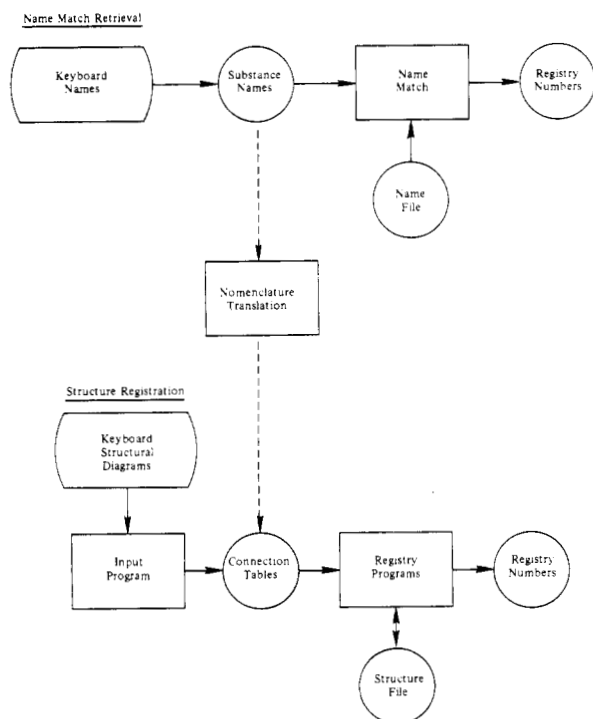


Figure 1. Retrieval from registry files.

Names. (This application is discussed more fully in the concluding section of this paper.)

BACKGROUND

The idea of automatic conversion of names to structure representations was discussed by Opler in 1958.⁶ Garfield described procedures for generating molecular formulas from systematic names;⁷ Rush and Elliott have discussed this possibility in terms of recent advances in grammatical analysis.⁸ Nomenclature translation has also been discussed by Dyson⁹ and by various workers in the USSR.¹⁰⁻¹³ Procedures for converting names of steroids to structural diagrams have recently been described by Stillwell.¹⁴

At CAS, work on nomenclature translation began with the preparation of an extensive algorithm for the conversion of systematic names to structural representations. The development of this algorithm, upon which the present program is based, was the subject of an earlier paper.¹⁵

PROCESSING A NAME BY NOMENCLATURE TRANSLATION

The basic input to the translation program should be an unambiguous, structure-based name and an associated identifying number. As shown in Figure 2, the input data may either be specifically keyboarded for translation or it may be derived by program from a computer-readable file of names, such as would be produced in normal CAS processing. The translation of a name is basically a matter of reading the name from left to right, recognizing and storing locant information[†] until needed, and matching the alphabetic strings with lists of alphabetic strings that are known to occur in particular roles in systematic names. As character strings are matched, the corresponding connection table fragments are set up, modified, and connected as indicated by the information in the name. (In a sense, the name itself is used as a program to direct the setting up of a particular

[†] Locants are the numeric, Greek, and italic characters that indicate the points at which structural fragments are connected together.

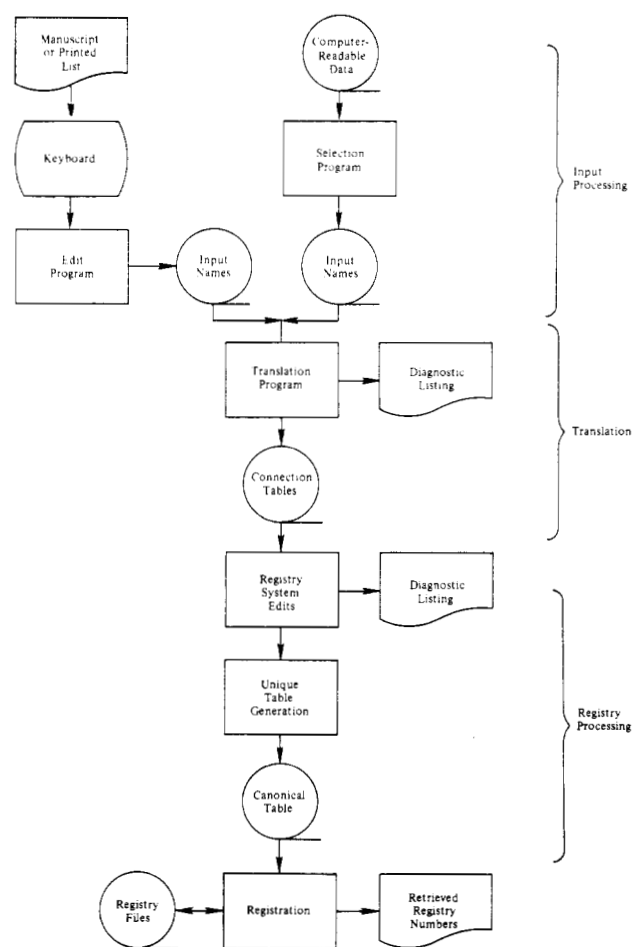


Figure 2. Processing names by nomenclature translation.

structure; by analogy, the translation program can be regarded as an unusual form of compiler program.) A successful translation produces a connection table which is then formatted for input to the CAS Registry System. If any information in the name is not recognized by the program or any part of the name is found to be ambiguous, impossible, or inconsistent, the translation is unsuccessful; the name is then rejected and an appropriate diagnostic is output.

The connection tables which are sent to the Registry System first pass through that system's editing program,¹⁶ which rejects tables inconsistent with their molecular formula or unacceptable in certain other respects. The tables which pass through the Registry edits are then converted into their canonical forms for matching against the structure file and retrieval of Registry Numbers (Figure 2).

The organization of the program can be illustrated in terms of its correspondence to the structure of CA Index nomenclature.¹⁷ A CA Index Name uses three segments to describe the atoms and bonds in a structure: Heading Parent, Substituent, and Name Modification. (Additional segments are used to give stereochemical information or to resolve certain types of synonymy.) The Heading Parent describes the main ring system or chain present and usually includes a suffix describing the main functional group. For example, in the name "4-Hexen-2-one, 3,3-diethyl-, hydrazone" the Heading Parent, "4-Hexen-2-one", describes the main chain, "hexen", and the functional group, "one". The Substituent describes radicals attached to the parent structure; in the example, this is "3,3-diethyl-". This portion is made from names of individual radicals, *e.g.*, "methoxy" and "phenyl". These radicals in turn are typically constructed from word roots such as "meth" and "phen",

AUTOMATED CONVERSION OF CHEMICAL SUBSTANCE NAMES

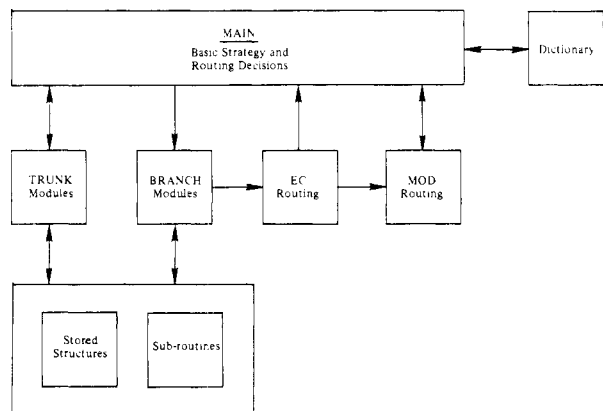


Figure 3. Organization of translation program.

which correspond to rings or chains, and suffixes such as "yl" or "oxy". The Name Modification—"hydrazone" in the example—describes either modifications to the primary functional group or molecular fragments which are not covalently bonded to the parent structure, as in the case of salts or addition complexes. Name Modifications are in part constructed from the same set of radical names used in Substituents.

Three portions of the translation program parallel these three segments of a name (see Figure 3). The TRUNK modules process the Heading Parent portion of the name. They identify the word root corresponding to the main ring or chain *via* dictionary look-up, retrieve the corresponding connection table fragment, analyze the remainder of the Heading Parent to identify recognized functional group suffixes, and modify the connection table accordingly. The BRANCH modules process the names of radicals used in the Substituent or Name Modification. The word roots describing the main ring or chain of each radical are identified by look-up in the same dictionary used by the TRUNK modules, but the program recognizes a different set of suffixes for a radical than for a Heading Parent based on the same word root. The MOD routine processes Name Modifications; it locates and changes the functional group modified by a term such as "ester" or "acetal". After the processing of each radical or modification, the EC (End of Component) subroutine uses the punctuation present to decide whether to assemble a complex radical (e.g., to attach three methoxy radicals to a benzene ring as in "(trimethoxyphenyl)", to proceed with identification of another radical, to transfer control to MOD, or to attach the held radicals to the parent structure. The MAIN routine makes the basic routing decisions that move control of the translation process from one area to another, and this routine also directs the organizing and formatting of output data.

Throughout the TRUNK, BRANCH, and MOD areas of the program, many functions are performed similarly regardless of where they occur. These functions are performed by the program's "subroutines", a group of about 40 routines which are called by various parts of the program as needed. (Program subroutines are described in Appendix I and are referred to by acronym in this discussion.) These subroutines include both high-frequency procedures, such as analysis of locant information or attachment of connection table fragments, and less frequently used, more specialized routines such as those which process spiro rings or place heteroatoms described by prefixes such as "oxa" or "aza".

The course of a typical name through these routines can be illustrated by the translation of an example

"4-Hexen-2-one, 3,3-diethyl-, hydrazone"

Table I. Connection Table for "Hex"

<div><div>C — C — C — C — C — C</div><div>6 5 4 3 2 1</div></div>			
Atom no.	Element	Locant names	Connections
1	C	1	— 2
2	C	2	— 1 — 3
3	C	3	— 2 — 4
4	C	4	— 3 — 5
5	C	5	— 4 — 6
6	C	6	— 5

Table II. Connection Table for "4-Hexen"

$\begin{array}{cccccc} \text{C} & - & \text{C} & = & \text{C} & - & \text{C} & - & \text{C} & - & \text{C} \\ 6 & & 5 & & 4 & & 3 & & 2 & & 1 \end{array}$			
Atom no.	Element	Locant names	Connections
1	C	1	- 2
2	C	2	- 1 - 3
3	C	3	- 2 - 4
4	C	4	- 3 = 5
5	C	5	= 4 - 6
6	C	6	- 5

Table III. Connection Table for "4-Hexen-2-one"

Atom no.	Element	Locant names	Connections
1	C	1	- 2
2	C	2	- 1 - 3 = 7
3	C	3	- 2 - 4
4	C	4	- 3 = 5
5	C	5	= 4 - 6
6	C	6	- 5
7	O		= 2

The major events in the translation of this name are as follows:

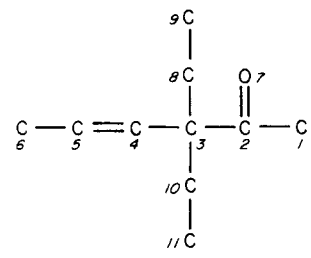
1. The locant "4" is recognized and held for later use.
2. The word root "hex" is identified by dictionary look-up. The corresponding connection table (Table I) is retrieved along with the identification of the appropriate TRUNK module to be used for analysis of the suffixes.[§]
3. The suffix "en" is recognized in the TRUNK module, and as a result the MLU subroutine places a double bond as directed by the held locant, "4". The result is shown in Table II.
4. The locant "2" is found and held. The suffix "one" is then recognized, resulting in an instruction to attach a double-bonded oxygen atom to the parent structure by means of the MLB subroutine. The result is shown in Table III. A record is also made which shows that atom 2 is the center of a "modifiable group", i.e., a group which may be modified by certain types of data in the Name Modification.
5. The punctuation after "one" indicates the end of the Heading Parent, and the MAIN routine causes analysis to be transferred to the BRANCH routines.
6. The locants "3,3" and the multiplying term "di" are recognized and held, the latter as its numeric equivalent "2". The word root "eth" is identified in the dictionary, and the appropriate connection table is retrieved. In the BRANCH routine which follows, the suffix "yl" is recog-

[§] The tables shown in these examples are representations of the connection tables being developed during translation. "Atom number" is the number by which an atom is identified in the connection table. "Element" is the element type of an atom. "Locant names" are labels which may be used in some part of a name to refer to a specific atom. In the "Connection" column, a "-" represents a single bond; a "=", a double bond. Thus, the listing for atom 2 in the table shown in this example "-1-3", means that atom 2 is connected to atoms 1 and 3 by single bonds.

Table IV. Connection Table for "ethyl"

Atom no.	Element	Locant names	Connections
1	C	1	-2
2	C	2	-1

Table V. Connection Table for "4-Hexen-2-one, 3,3-diethyl-"



Atom no.	Element	Locant names	Connections
1	C	1	-2
2	C	2	-1 -3 =7
3	C	3	-2 -4 -8 -10
4	C	4	-3 =5
5	C	5	=4 -6
6	C	6	-5
7	O		=2
8	C		-3 -9
9	C		-8
10	C		-3 -11
11	C		-10

nized. Atom 1 of this structure is labeled as a "point of attachment", *i.e.*, the atom to be used in attaching this structure to the parent. The connection table for the radical is held along with the locants and multiplier (Table IV).

7. The EC subroutine finds the "-", indicating the end of the Substituent, in this case "3,3-diethyl". The held radical is then attached to the parent using the MLB subroutine, which makes use of the stored locant and multiplier information and the point of attachment. The combined table is shown as Table V.

8. The word root "hydraz" is identified in the dictionary, and the suffix "one" is found in the corresponding BRANCH routine. In processing the "hydrazone" modification, the structure already set up is examined to find the appropriate type of modifiable group (atoms 2 and 7 in this case) and the "hydraz" connection table (for the structure N-N) is used to replace the oxygen atom. See Table VI.

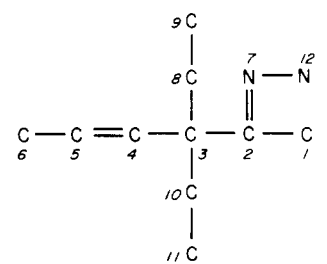
9. The end of the name is recognized, and the MAIN routine causes the connection table to be formatted for output.

PROCESSING SPECIAL CHARACTERISTICS OF NOMENCLATURE

The previous example represents a relatively straightforward translation; its most subtle point is the simple modification "hydrazone". However, to increase the nomenclature translation program's applicability to systematic nomenclature, it has been necessary to develop special procedures for handling many other aspects of nomenclature. Some of these are discussed below.

Complex Radicals. Processing a compound radical name such as "(4-methoxyphenyl)" or "(dimethylamino)" requires that the connection tables corresponding to individual radical names be held with their locants and multipliers until the "parent" of the compound radical ("phenyl" and "amino" in the two examples) is identified. The

Table VI. Connection Table for "4-Hexen-2-one, 3,3-diethyl-, hydrazone"



Atom no.	Element	Locant names	Connections
1	C	1	-2
2	C	2	-1 -3 =7
3	C	3	-2 -4 -8 -10
4	C	4	-3 =5
5	C	5	=4 -6
6	C	6	-5
7	N		=2 -12
8	C		-3 -9
9	C		-8
10	C		-3 -11
11	C		-10
12	N		-7

entire radical is then assembled and held until it can be attached to the parent. Often this analysis must proceed through several levels, as in a name such as "2-Isoindolinopropionamide, N-[2-[[[2-(dimethylcarbamoyl)ethyl]methyl]carbamoyl]ethyl]-N-methyl-1,3-dioxo-", where six different levels are involved. The locants and multipliers at each level must be correctly associated with the corresponding structural fragments as indicated by the brackets and parentheses in the name.

Omission of Understood Locants. For "parent" radicals such as "methyl" or "amino", where only one atom can receive attachments, locants are not required to indicate where attachments are to be made. Similarly, locants are omitted where equivalent atoms are available for substitution, as in the name "cyclohexane, methyl-". In the translation program such names are handled by marking atoms which may receive such attachments as "attachment acceptors"; this indicates that a marked atom may receive a specified number of attachments without requiring locants. If locants are omitted and no such marked atoms are present, the name will be rejected as ambiguous, as in the case of "2-butanol, chloro-". An exception is made in a case where all available positions are to be substituted, as with "ethanol, pentafluoro-".

Stereochemistry. Within the CAS Chemical Registry System, the main part of the atom-bond connection table describes the two-dimensional graphical representation of a substance. An additional field, called the Text Descriptor, resolves cases where two or more different substances have identical two-dimensional graphs. This field, an alphanumeric description of the stereochemistry of a substance, closely parallels the terms and special characters used to describe stereochemistry in systematic nomenclature. During processing by the translation program, items of stereochemical data are identified at various points in TRUNK and BRANCH processing. After translation is completed these items are ordered and formatted in a descriptor acceptable to the Registry programs.

Conjunctive Names. The term "conjunctive name" refers to a name formed by concatenation of names which may be used independently. For example, "cyclohexane" and "ethanol" may be used to form the conjunctive name "cyclohexaneethanol". Such names are frequently used as Heading Parents in CA Index Names in order to bring sub-

Example: 1,2-Benzenediethanol

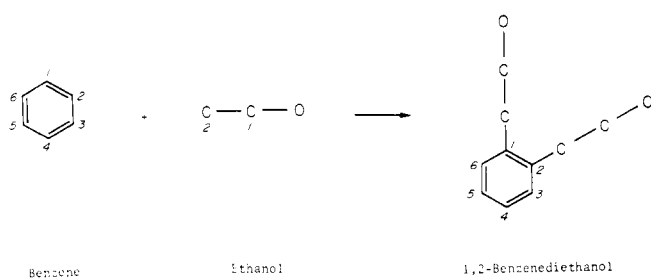


Figure 4. Processing a conjunctive name.

stances based on the same ring system together in the Index. When such a name is translated, the first parent name ("cyclohexane" in this case) is processed normally, but the characters following it do not correspond to any of the allowed functional group suffixes. A special switch is then set and the remainder of the Heading Parent is processed as an independent parent name. If the second name can be fully processed and the "conjunctive" switch is set, the second structure is attached to the first as if it were an ordinary functional group. An example is shown in Figure 4.

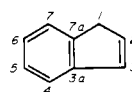
Bridged Rings Named by the von Baeyer System. For ring system names such as "bicyclo[4.2.0]octane" or "tetracyclo[2.2.1.0^{2,6}.0^{3,5}]heptane", the connection table of the ring system is fully described by the numbers contained in brackets and is set up by an algorithm which uses these numbers. The word root which follows the brackets, *e.g.*, "oct" or "hept", makes possible a consistency check to determine whether the number of atoms described by the bracketed numerics equals that described by the word root.

Spiro Rings. Ring systems with spiro junctions are named either as simple spiro systems, such as "spiro[2.5]octane", or complex spiro systems, such as "spiro[indene-1,1'(2'H)-naphthalene]". Names of the first type are handled in a manner similar to that used for the bridged systems described above. In processing complex spiro names, the identification of "spiro" followed by another name results in the setting of a switch which governs handling of these ring system names: the first system is identified and its connection table is set aside until the second has been processed; the two systems are then joined with loss of one atom and adjustment in the locant names of the atoms involved (see Figure 5).

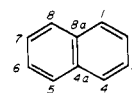
Fused Ring Systems. Many fused ring systems have names in which the fusion faces are described by alphabetic or numeric characters in brackets, *e.g.*, "thieno[2,3-*d*]pyridine" or "2*H*, 11*H*-tripyrido[1,2-*a*:1',2'-*c*:3'',2''-*e*]pyrimidine". Names of this type cannot be processed by a simple algorithmic treatment like that used for bridged systems because of difficulties in establishing the correct enumeration of the atoms in more complex systems. The rules for enumeration depend on finding the correct orientation of a ring system on paper, a process not readily described by algorithm. Therefore, these names are handled through the use of a special listing of known fusion names associated with their corresponding connection tables. When a name involving such a ring system is recognized, the name of the ring system is isolated and identified on this special list and the appropriate connection table is retrieved. The remainder of the translation proceeds normally.

Omission of Enclosing Marks. Nomenclature practice in CA Indexes has changed over the years, particularly as less systematic names, such as "enanthic acid" and "testosterone", which were used in earlier Collective Index periods, were eliminated from the vocabulary of CA Index nomenclature. (Most such names were eliminated at the be-

Indene



Naphthalene



Atom Number	Element	Locant Name	Connections	Atom Number	Element	Locant Name	Connections
1	C	1	-2-9	1	C	1	=2-10
2	C	2	-1-3	2	C	2	=1-3
3	C	3	=2-8	3	C	3	=2-4
4	C	4	=8-5	4	C	4	=3-9
5	C	5	=4-6	5	C	5	=6-9
6	C	6	=5-7	6	C	6	=5-7
7	C	7	=6-9	7	C	7	=6-8
8	C	3a	-3-4=9	8	C	8	=7-10
9	C	7a	-1-7=8	9	C	4a	-4-5=10
				10	C	8a	-1-8=9

Spiro[indene-1,1'(2'H)-naphthalene]

Atom Number	Element	Locant Name	Connections
1	C	1	-2-9-10-18
2	C	2	-1-3
3	C	3	=2-8
4	C	4	=5-8
5	C	5	=4-6
6	C	6	=5-7
7	C	7	=6-9
8	C	3a	-3-4=9
9	C	7a	-1-7=8
10	C	2'	-1-11
11	C	3'	-10=12
12	C	4'	=11-17
13	C	5'	=14-17
14	C	6'	=13-15
15	C	7'	=14-16
16	C	8'	=15-18
17	C	4'a	-12-13=18
18	C	8'a	-1-16=17

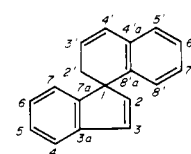


Figure 5. Processing a complex spiro name.

ginning of the 9th Collective Index period with CA Volume 76.) In general, names which are no longer preferred can be handled within the framework of the translation program by including the outdated word roots in the dictionary. One major variation in punctuation practice in earlier volumes, however, has required special treatment. Prior to CA Volume 56, complex radicals such as "2-hydroxyethyl" were not placed within enclosing marks when the names of the component radicals were not in alphabetic order. For example, "Hexanoic acid, 2-(methylamino)-" in earlier indexes is described as "Hexanoic acid, 2-methylamino-". These names are still specific and unambiguous so long as the conventions used are understood. To handle these names, a routine was prepared which checks whether the individual radicals are in proper alphabetic sequence; when they are not (and a name from an earlier volume is being treated), the routine causes the name to be correctly interpreted. For more recent names, this same technique can be used to detect errors in alphabetization of radicals.

PROGRAM SIZE AND TIMINGS

The nomenclature translation program is written in 360/370 Assembler language for use on the IBM 360/370 series or compatible equipment.* The total program at present requires about 205,000 (205K) bytes of machine storage. At CAS the program is normally run in a 200K partition, with a few infrequently used routines kept on disk until they are required. Using a smaller partition, with greater amounts of the program kept on disk, appreciably increases machine time because of the overhead costs of additional moving of programs from disk to core. Since any given name requires only a few portions of the total program, the program operates quite rapidly if the needed routines are available in core. The program in a 200K partition

* Chemical Abstracts Service does not endorse specific items of hardware or software. Use by Chemical Abstracts Service, therefore, implies no product endorsement.

processes approximately 4900 names per minute on the IBM 370/165, including both successful translations and rejections; i.e., about 0.01 sec is required per name. With an 88K partition, the time required is increased by about a factor of 10.

Since the dictionary of word roots must usually be accessed several times during the processing of a name, the efficiency of dictionary lookup is critical to the overall program efficiency. Therefore, the dictionary of word roots is kept in core regardless of the partition size. Because of this, the dictionary has been kept relatively small. At present it contains 321 terms. While most of the word roots in the dictionary represent single ring systems or chains, some smaller and larger terms have been included to handle particular areas of nomenclature efficiently. For example, most word roots in the dictionary do not include suffixes, but in some cases, as with the phosphorus acids, inclusion of partial suffixes appears to be the more practical route. Because of the frequency of names beginning with "benz" and because many fused ring names begin with "benz" (or "benzo"), special treatment has been given to the family of terms beginning with "benz". The identification of "benz" leads to a 37-item second-level dictionary. The dictionary is searched by a subroutine called SMC, which determines whether any dictionary entries have the same first three letters as the candidate term. If the first three letters are matched (there are 174 different three-letter sequences at present), SMC attempts to complete a match of the candidate term with one of the dictionary terms beginning with the three letters. This strategy increases the efficiency of dictionary look-up by deferring the problem of matching variable-length word roots until there is a high probability that a match will occur.

CURRENT PROGRAM CAPABILITIES

The nomenclature translation program has been tested on several thousand names of organic (i.e., carbon-containing) compounds. This testing has included names chosen specifically to test particular points in the program and has also involved representative samples of names chosen to provide data on program performance and to indicate which types of nomenclature can be handled by the program at this time and which types cannot. Successful translations have been validated either by examining printouts of the resulting connection tables or, where Registry Numbers for the substances were known, by using the generated connection tables to retrieve Registry Numbers from CAS Registry System files (see Figure 3). In cases where the expected number was not retrieved, a determination was made of whether the error was in the name, in the structure files, or in the program logic or coding. The tests discussed here included both samples of CA Index Names and samples of names from primary documents.

CA Index Names. During a period of several months in late 1972 and early 1973, all new CA Index Names were processed through the nomenclature translation program. The results were studied as a means of evaluating the program's effectiveness as a tool for editing CA Index Names. The current program was found to be capable of successfully translating about 45% of the new CA Index Names, not counting names which contained errors. (A cross section of correctly translated names is shown in Appendix II.)

The names that were not translated successfully can be divided into three major classes: (1) names which contain inadequate, ambiguous, or incorrect data; (2) non-structure-based names such as laboratory numbers and *Colour Index* names; and (3) names which are clear and unambiguous but which contain terms, suffixes, or conventions not recognized by the program.

Most names which are in error fall into the first class. However, if the error causes the name to describe a valid

structure the name is not rejected. For example, if the name intended was "1-pentanol" and, by mistake, "2-pentanol" was keyboarded, the name would not be rejected. The program would, however, reject the misspelling "1-Pentanol", the misformatted "1Pentanol", or the impossible "11-Pentanol". This class of rejections also includes generic names such as "butanol, chloro-", which do not distinguish among the possible isomers. These and any other names which the program determines to be ambiguous will be rejected.

Only 5-10% of CA Index Names of unique substances belong to the second class. These names are completely unrelated to the structure being described and thus are not amenable to nomenclature translation techniques.

Names which are systematically constructed and unambiguous but not handled by the program contain one or more features—word roots, suffixes, or other conventions—not anticipated by the program.[†] For experimental development of the program, only features which occurred relatively frequently in CA Index nomenclature were chosen for inclusion. Most CA Index Names that are not currently handled by the program could be translated if the program's dictionary, suffix lists, etc., were sufficiently large.

On the basis of these data, a number of modifications and extensions have been defined which will make the translation program's coverage of current CA Index Names much more nearly complete. An upgraded version of the program which will provide a more effective name editing tool is being prepared and should be operational during 1975.

Names from Primary Documents. Almost all of the names discussed thus far have appeared in the "inverted" form used in CA nomenclature. In an "inverted" name, the Heading Parent segment is placed ahead of the Substituent and Name Modification segments. The use of inverted names causes names having the same Parent name to be brought together in the Index. For example, the inverted names "2-butanol, 1-chloro-" and "2-butanol, 1-iodo-" will appear close together, whereas the corresponding "uninverted" names "1-chloro-2-butanol" and "1-iodo-2-butanol" would be widely separated.

Substance names used in primary documents are normally not inverted because listing is usually not important in a primary publication. (CA Indexes have also used uninverted names for ionic organic compounds—"onium" names—indexed prior to 1967). Techniques for processing uninverted names have been incorporated into the translation program. If an associated code indicates to the program that a name is not inverted, processing begins in the BRANCH routines. When a potential failure point is reached, the INVERT subroutine is used to generate an inverted form of the name. If INVERT succeeds, the inverted name is processed by the program in place of the uninverted name. Figure 6 shows some uninverted names and the corresponding "intermediate" inverted names which were generated and correctly translated.

While no completely representative sample of primary-literature nomenclature has been derived and tested, a reasonable picture can be obtained from a sample that was prepared from two journals—the *Journal of Organic Chemistry* and the *Journal of the Chemical Society C*. During several weeks in 1970, the names of all substances chosen as index entries from these journals were input to the translation program. Of these names, 41% were successfully translated; some examples are given in Appendix III. Among the reasons for rejections are those described above for Index Names. In addition, some names follow punctuation and spelling conventions which are different from those of CA nomenclature. Since the program anticipates

[†] These features include the names of coordination compounds, about 10% of the substances indexed from CA.

1. Uninverted name:
1-fluoro-1,1-dinitro-2-propanol
▲
Inverted name:
2-propanol, 1-fluoro-1,1-dinitro-
2. Uninverted name:
1-(allyloxy)-4-nitrobenzene
▲
Inverted name:
benzene, 1-(allyloxy)-4-nitro-
3. Uninverted name:
ethyl 4,4,4-trichlorobutyrate
▲
Inverted name:
butyric acid, 4,4,4-trichloro-, ethyl ester
4. Uninverted name:
tris(2-chloroethyl)sulfonium chloride
▲
Inverted name:
sulfonium, tris(2-chloroethyl)-, chloride

('▲' indicates point of potential rejection.)

Figure 6. Processing uninverted names.

strict adherence to these conventions, variations are likely to lead to rejections.

NOMENCLATURE TRANSLATION FOR NAME EDITING

Although CAS has eliminated many of the exceptions and irregularities formerly found in the CA Index Name selection rules, the application of these rules continues to require highly trained nomenclature specialists. As the number of substances per volume has continued to grow, economic considerations make it necessary to reduce the amount of work which must be done by these specialists both in preparing names and in verifying and editing them. The work required for editing new names can be reduced by using computer-based editing procedures to identify those CA Index Names which contain questionable or inconsistent data. The nomenclature specialists need then review only the names that are questioned instead of verifying every name in detail as is required in a completely manual system.

An early CAS program to perform some edits on chemical names was described by Park.¹⁸ Subsequent versions of this program have incorporated a number of additional editing features which are now applied to all new chemical names being entered into CAS files.

The nomenclature translation process offers a potential means of performing much more extensive editing of CA Index Names. Since a name must be consistent and unambiguous in order to be translated, any name which does not meet these criteria will be rejected, and an appropriate diagnostic message will be indicated to aid in review of the name. The thorough name analysis required for translation provides an opportunity for other edits which are not inherently part of translation but require a similar analysis. These include, for example, checks on whether names of radicals are correctly alphabetized or on whether the lowest possible numbering has been chosen for locants which describe attachments to a symmetrical structure. These edits would be prohibitively expensive as isolated steps, but their costs will most likely be manageable if they are embedded in the translation process.

A name editing system based on nomenclature translation is currently being implemented for use in CAS indexing operations. This system should be capable of detecting

most errors in new CA Index Names. The system will have three stages at which error detection can occur: (1) the preliminary editing program; (2) nomenclature translation itself, with a name being either converted to a connection table or rejected; (3) comparison of structure representations generated from names with structure representations already on file in the CAS Chemical Registry System. If the two structural representations are the same, the name and the structure for a substance can be assumed to be consistent; if not, the discrepancy must be reviewed and any errors in the name or structure must be corrected.

ACKNOWLEDGMENTS

The financial support provided by the National Science Foundation is gratefully acknowledged, as are the important contributions made to the programming effort by Mr. Pennell Watkins, Mr. Anthony Thorp, and Mrs. Jeri Cowan.

APPENDIX I NOMENCLATURE TRANSLATION SUBROUTINES

- ACT (Attach Connection Tables)—Connects one structure to another at designated positions.
- ADDAT (Add Atom)—Adds single atom to work area connection table.
- ADIB—Handles assembly of complex doubling radicals.
- AIR (Add Information Record)—Relates secondary structural information to designated atom in connection table.
- AL (Alphabetic)—Determines whether radicals are out of alphabetic order, and, for pre-1962 Index Names, directs assembly of complex radicals which appear without enclosing marks.
- ANDE (Anhydro and Deoxy)—Performs removal of oxygen atoms as indicated by radicals "anhydro" and "deoxy".
- BI—Accomplishes doubling of parent structure as indicated by prefix "bi".
- BR (Bridges)—Directs placement of bridging structures which have been named by appropriate prefixes.
- CBT (Change Bond Type)—Changes specified bond in structure to designated bond type.
- CCT (Copy Connection Table)—Sets up identical copy of given structure.
- CET (Change Element Type)—Replaces specified atom in connection table with designated element.
- CJ (Conjunctive Names)—Assembles structures named by conventions of conjunctive names or structures named using the terms "compd. with" or "ester with".
- CSK (Cancel Skeleton)—Deletes structure which is no longer needed.
- CYC (Cyclic)—Builds cyclic structures named by von Baeyer or simple spiro names.
- DCT (Double Connection Table)—Prepares copy of given connection table and assigns primed locants to atoms in copy.
- DME (Determine Multiplier Equivalent)—Determines binary number equivalent to given multiplier word root.
- EC (End of Component)—Makes routing decisions based on punctuation found at the end of given radical name.
- FLR (Find Locant Rank)—Searches locant name list to find corresponding connection table rank number.
- FRH (Find Rank of Highest)—Searches locant name list for highest numbered locant.
- FUS (Fused)—Determines presence of ring fusion term and directs analysis to special dictionary of fused ring names.
- HAL (Halogens)—Identifies halide term and designates corresponding atoms.

HM (Hold Multiplier)—Associates binary equivalent of multiplier with proper locants and structure.

HOLD—Performs functions necessary to hold various types of data with associated locants and multipliers.

HST (Hold Stereo)—Holds stereochemical terms for later reordering into text descriptor.

HYD (Hydro)—Changes ring double bonds to single bonds as directed by "hydro" term with appropriate locants and multipliers.

LMG (Locate Modifiable Groups)—Finds groups which are to be changed as directed by Name Modification term or by ANDE or THI.

LOC (Locant)—Searches for, identifies, and holds locants and indicated hydrogen terms.

MLB (Multiplier-Locant-Branch)—Attaches held branch or subbranch structures to parent or main branch, using appropriate locants, multipliers, indicated hydrogens, points of attachment, and attachment acceptors.

MLH (Multiplier-Locant-Hetero)—Directs the replacement of carbon atoms by heteroatoms named by prefixes, using appropriate locants and multipliers.

MLU (Multiplier-Locant-Unsaturation)—Changes single bonds to double or triple as directed by suffixes "en" or "yn", using appropriate locants and multipliers.

N—Attempts to unambiguously assign locants such as *N*, *N'*, and *N^x* to nitrogen atoms in a structure.

PATH—Determines if two designated atoms are joined by alternating bond system or by chain or specified length.

PLC (Primed Locant Control)—Determines number of primes to be added to locant names when copying structure.

RAC (Remove Atom and Connections)—Removes all reference to given atom from connection table.

RAG (Remove Arabic or Greek)—Removes all Arabic or Greek locant names from locant name list, or moves all Greek locants to next higher rank number.

SCT (Set Up Connection Table)—Expands compact connection table associated with the dictionary to redundant connection table and sets up the corresponding index and information lists.

SMC (String Match Completion)—Determines whether alphabetic string can be matched in dictionary of variable length word roots; if matched, retrieves corresponding connection table and direction to routine to be used for analysis of suffixes.

SP (Spiro)—Joins ring systems as designated by spiro name.

ST (Stereo)—Organizes held stereochemical terms into text descriptor ordered and formatted as required by CAS Registry System.

THI (Thio)—Replaces oxygen by sulfur as directed by "thio" term, using appropriate locants and multipliers.

APPENDIX II REPRESENTATIVE SUCCESSFULLY TRANSLATED CA INDEX NAMES

Phosphinic amide, *P,P*-diphenyl-*N*-2-thiazolyl-
Silane, [2-(3-cyclohexen-1-yl)ethyl]triethoxy-
1,2-Cyclohexanediamine, *trans*-
Phosphorothioic acid, *O,O*-dimethyl *S*-(2-nitrophenyl) ester
Piperidine, 2,6-dimethyl-, *trans*-
Naphthalene, decahydro-1-methyl-, (1 α ,4 α ,8 α)-
Phosphorodiamidic acid, *N,N*-bis(2-chloroethyl)-, phenyl ester
Benzeneethanamine, *N,N*-bis(2-methylpropyl)-
Benzenebutanoic acid, α -propyl-
2*H*-Quinolizin-2-ol, octahydro-, *trans*-
1*H*-Benzimidazole-4,7-dione, 5-bromo-6-(methylamino)-2-(trifluoromethyl)-
Benzenesulfonic acid, 4-(acetylamino)-2-amino-
2-Butenenitrile, 3-bromo-, (*Z*)-
Benzenamine, *N,N*-dimethyl-3-(methylthio)-
4(3*H*)-Quinazolinone, 2-(chloromethyl)-3-(phenylmethoxy)-

Pregna-4,6-diene-3,20-dione, 21-(acetyloxy)-6-fluoro-17-hydroxy-16-methylene-
Propanedioic acid, [(2,4-dichlorophenyl)methylene]-
Pregn-5-ene-12,20-dione, 3,14,15-trihydroxy-, (3 β ,14 β ,15 α)-
Phosphorane-carboxylic acid, dichlorodimethyl-ethyl ester
2(3*H*)-Benzothiazolethione, 3-[(acetyloxy)methyl]-5-chloro-
Benzo[*b*]thiophen-3(2*H*)-one, 2-(phenylmethylene)-, 1,1-dioxide
2*H*-1,2-Benzothiazin-3(4*H*)-one, 2-butyl-, 1,1-dioxide
9*H*-Fluorene-9-carboxylic acid, 3-methyl-, methyl ester
3(4*H*)-Quinazolinepropanaminium, *N,N,N*-triethyl-6-iodo-2-methyl-4-oxo-, iodide
1-Hexanesulfonic acid, 2-ethyl-, sodium salt
5,9-Undecadien-2-one, 10-(5,5-dimethyl-1,3-dioxan-2-yl)-6-methyl-
D-Glucose, 3-*O*-[2-(acetylamino)-2-deoxy- α -D-galactopyranosyl]-
3-Hexanone, 6-[2-ethyl-5-(1-ethyl-2-oxobutyl)tetrahydro-4-methyl-2-furanyl]-6-hydroxy-
Pyridazine, 3-chloro-6-(methylsulfinyl)-, 2-oxide
Bicyclo[3.1.0]hexan-3-ol, 6,6-dichloro-, (1 α ,3 β ,5 α)-
1,2-Ethanediamine, *N'*-[2-(diphenylphosphino)-1-methethyl]-*N,N*-diethyl-
3-Benzofurancarboxylic acid, 5-methoxy-2-[(4-methyl-1-piperazinyl)methyl]ethyl ester, dihydrochloride
Hydrazinecarbothioamide, 2-[1-(4-chlorophenyl)ethylidene]-*N*-(4-methoxyphenyl)-
2-Undecenoic acid, 7-oxo-, (*E*)-
Ethanethioic acid, 5-(1*H*-inden-3-ylmethyl) ester
Benzo[*b*]thiophene-6-carboxylic acid, 4-methoxy-7-methyl-
4*H*-1-Benzopyran-4-one, 3-[(dimethylamino)methyl]-2,3-dihydro-, hydrochloride
9,10-Anthracenedione, 1-amino-2-bromo-5,8-dihydroxy-4-[(4-methylphenyl)amino]-
Phenol, 4-chloro-2-[5-(4-chlorophenyl)-4,5-dihydro-1-phenyl-1*H*-pyrazol-3-yl]-5-methyl-
Benzoic acid, 4-[(4-azidophenyl)azo]-, methyl ester
3-Thiophenemethanol, 5-(4-chlorophenyl)-4-methoxy-
3-Thiopheneacetic acid, 4-methoxy- α -methyl-5-phenyl-, sodium salt, (+)-
3-Thiopheneacetic acid, 5-(3-chlorophenyl)-4-methyl-, ethyl ester
Phosphine oxide, decylidenebis(dimethyl-
Benzenemethanol, α -(aminomethyl)-3,4-dimethoxy-, hydrochloride, (*R*)-
Glycine, *N*-[1-[*N*-[(4-bromophenyl)methoxy]carbonyl]glycyl]-L-prolyl]-L-leucyl]-
Cholestane-4,5-diol, 3-methoxy-, (3 β ,4 α ,5 α)-
3-Pyridazinethiol, 6-ethoxy-, 2-oxide, sodium salt
Silanamine, 1,1,1-trimethyl-*N*-(2-methyl-1-propenyl)-*N*-propyl-
Pyrimidine, 4,6-dichloro-2,5-dihydro-5-methyl-2-(1-methylbutylidene)-5-propyl-

APPENDIX III REPRESENTATIVE SUCCESSFULLY TRANSLATED NAMES TAKEN FROM PRIMARY JOURNALS

The following are a cross section of the names that were taken from several months of the *Journal of Organic Chemistry* and the *Journal of the Chemical Society, C* and successfully translated.

2-(1,1-Dimethyl-2-propynyl)cyclopentanol
o-Isopropylphenyl *m*-nitrobenzenesulfonate
Diethyl [[2-(ethoxycarbonyl)-4,5-dimethylpyrrol-3-yl]methyl]-malonate
3-Isobutoxy-1-propanethiol
1-[(5-Carboxy-3,4-dichloropyrrol-2-yl)methyl]pyridinium bromide ethyl ester
2-Fluoro-2,2-dinitroethanol
N-(1-Methylheptyl)benzamide
(*E*)-2',4',4',6'-Tetrahydroxy-3-methoxychalcone
[3-(Diisopropylamino)-2-propynyl]trimethylsilane
3-Methyl-2-cyclopenten-2-ol-1-one
Cycloheptanone dimethyl ketal
3 β -Nitrosocholestane

1,2,3,4-Tetrahydro-1-oxo-2-naphthalenepropionamide
 Methyl (*E*)-3-(cyclohexylamino)acrylate
 3-Acetoxy-1,1-dimethylcyclohexane
 6,6,11,11,18,18,23,23-Octamethyloctacosane
 2-Phenyl-6-*tert*-butyl-8-methylchroman
 1-Ethyl *N*,2-diphenylaspartate
 2,2-Dimethyl-3-nonanol
 3,4-Epoxy-3-methyl-4-phenylbutyric acid sodium salt
 Bis(3-phthalidyl)nitromethane
 1-Phenyl-3-*p*-tolyl-1,3-propanedione
 2'-Benzoyloxy-4,4'-dimethoxychalcone
 1-Benzyl-4-hydroxy-5-oxo-3-pyrroline-3-carboxylic acid *tert*-butyl ester
 2,3-Dimethyl-2-butenenitrile
 1,6-Dibenzyl-2,5-diphenyl-1,6-dihydropyrazine
N-Bromo-*N*-methyl-4-nitrobenzamide
 2-(1-Naphthyl)ethanol
 7-Methoxy-5-phenyl-1,3,4,5-tetrahydro-2*H*-1,4-benzodiazepin-2-one
 Terephthalic acid bis[*p*-(ethoxycarbonyl)phenyl] ester
p-(2,2,2-Trifluoroacetamido)benzoic acid ethyl ester
 2,4-Dimethyl-1-phenyl-3-pentanone
 6-(Dimethylamino)-3-pyridinecarboxylic acid methyl ester
 1-(*p*-Methylphenyl)-2-bromopropane
 Adamantan-2-one
 6-*tert*-Butyl-1,4-dimethylnaphthalene
 2-Phenylacetophenone oxime
p-Nitrophenol
 (2-Acetyl-1,4-dioxo-4-indanyl)trimethylammonium perchlorate
 α -Ethoxyethyl cyclopropyl ketone semicarbazone
 Diethylphosphine
N-Methylphthalimide
 7-Methylene-5 α -pregnan-20-one
 4-Amidino-1,3-dimethyl-4-phenylpiperidine dihydrochloride
 4,4'-Dimethoxybenzophenone
 3-Methylcrotonaldehyde
 Tetrahydropyran-3-ol
 1,3-Diphenylpyrazole
 1,2,3,4-Tetrahydro-4-methylcinnoline
 1-Benzyl-3-phenylindole

LITERATURE CITED

- Leiter, D. P., Jr., Morgan, H. L., and Stobaugh, R. E., "Installation and Operation of a Registry for Chemical Compounds," *J. Chem. Doc.*, **5** (4), 238-42 (1965).
- Farmer, N. A., Tate, F. A., Watson, C. E., and Wilson, G. A., "Extension and Use of the CAS Chemical Registry System," CAS Report No. 2, 3-10, April 1973.
- Rowlett, R. J., Jr., and Tate, F. A., "A Computer-Based System for Handling Chemical Nomenclature and Structural Representations," *J. Chem. Doc.*, **12** (2), 125-8 (1972).
- Conrow, K., "Computer Generation of Baeyer System Names of Saturated, Bridged, Bicyclic, Tricyclic, and Tetracyclic Hydrocarbons," *J. Chem. Doc.*, **6** (4), 206-212 (1966).
- Van Binnendyk, D., and Mackay, A. C., "Computer-Assisted Generation of IUPAC Names of Polycyclic Bridged Ring Systems," *Can. J. Chem.*, **51** (5), 718-723 (1973).
- Opler, A., "On the Automatic Manipulation of Representation of Chemical Structures," *Amer. Doc.*, **10**, 59 (1958).
- Garfield, E., "An Algorithm for Translating Chemical Names to Molecular Formulas," *J. Chem. Doc.*, **2** (3), 177-9 (1962).
- Elliott, P. M., and Rush, J. E., "Translation of Chemical Nomenclature by Syntax-Controlled Techniques," presented to the 6th Middle Atlantic Regional Meeting of the American Chemical Society, Baltimore, Md., Feb 1971.
- Dyson, G. M., "A Cluster of Algorithms Relating the Nomenclature of Organic Compounds to their Structure Matrices and Ciphers," *Inform. Storage Retrieval*, **2**, 59 (1964).
- Tsukerman, A. M., and Terentiev, A. P., "Chemical Nomenclature Translation," "Proceedings of the International Conference on Standards for a Common Language for Machine Searching and Translation," Vol. I, Interscience, New York, N. Y., 1960, p 493.
- Seifer, A. L., and Shtein, V. S., "An Algorithm for Conversion of a Name of a Complex Compound, Given in a Rational Nomenclature, into a Linear Formula," *Nauch.-Tekh. Inform., Vses. Inst. Nauch. Tekh. Inform.*, No. 1, 172 (1960).
- Stetsyura, G. G., and Tsukerman, A. M., "An Automatic Translation of Names of Organic Compounds into Formulas," *Nauch.-Tekh. Inform., Vses. Inst. Nauch. Tekh. Inform.*, No. 3, 17-19 (1962).
- Tsukerman, A. M., "Algorithm for Name Translation of Condensed Polycyclic Structures," *Nauch.-Tekh. Inform., Vses. Inst. Nauch. Tekh. Inform.*, No. 4, 23-30 (1965).
- Stilwell, R. N., "Computer Translation of Systematic Chemical Nomenclature to Structural Formulas-Steroids," *J. Chem. Doc.*, **13** (3), 107-9 (1973).
- Vander Stouw, G. G., Naznitsky, I., and Rush, J. E., "Procedures for Converting Systematic Names of Organic Compounds into Atom-Bond Connection Tables," *J. Chem. Doc.*, **7** (3), 165-9 (1967).
- Leiter, D. P., Jr., and Morgan, H. L., "Quality Control and Auditing Procedures in the Chemical Abstracts Service Compound Registry," *J. Chem. Doc.*, **6** (4), 226-9 (1966).
- Donaldson, N., Powell, W. H., Rowlett, R. J., Jr., White, R. W., and Yorka, K. V., "Chemical Abstracts Index Names for Chemical Substances in the Ninth Collective Period (1972-1976)," *J. Chem. Doc.*, **14** (1), 3-15 (1974).
- Park, M. K., Kenny, J. K., and Swartzentruber, P. E., "Automatic Editing of Chemical Nomenclature," presented at the 154th National Meeting of the American Chemical Society, Chicago, Ill., Sept 1967.