(7) a. Feldman, A., Holland, D. B., and Jacobus, D. P., J. CHEM. Doc. **3**, 187 (1963).
b. Mullen, J. M., *ibid.*, **7**, 88 (1967).

(8) a. Cossum, W. E., Krakiwsky, M. L., and Lynch, M. F., J. CHEM. Doc. **5**, 33 (1965).
b. Sussenguth, E. H., Jr., *ibid.*, **5**, 36 (1965).

(9) a. Jaffe, H. H., and Orchin, M., "Symmetry in Chemistry," John Wiley and Sons, Inc., New York, 1965, Chapters 1–3.
b. Mislow, K., "Introduction to Stereochemistry," W. A. Benjamin, Inc., New York, 1965, Chapter 1.

(10) a. McDonnell, P. M., and Pasternack, R. F., J. CHEM. Doc. **5**, 56 (1965).
b. Pasternack, R. F., and McDonnell, P. M., *Inorg. Chem.* **4**, 600 (1965).

(11) Cahn, R. S., Ingold, C. K., and Prelog, V., *Angew. Chem. Intern. Ed. Engl.* **5**, 385 (1966).

(12) Hohn, F. E., "Elementary Matrix Algebra," 2nd ed., The MacMillan Co., New York, 1965, pp. 42–45.

(13) McCasland, G. E., "A New General Method for the Naming of Stereoisomers," Chemical Abstracts Service, Columbus, Ohio, 1953, p. 2.

(14) Noller, C. R., "Chemistry of Organic Compounds," 2nd ed., W. B. Saunders Co., Philadelphia, 1957, pp. 383–84.

(15) Leiter, D. P., Jr., Morgan, H. L., and Stobaugh, R. E., J. CHEM. Doc. **5**, 238 (1965).

(16) a. See *(11)* p. 388.
b. Cahn, R. S., *J. Chem. Educ.* **41**, 116 (1964).

Procedures for Converting Systematic Names of Organic Compounds into Atom-Bond Connection Tables*

G. G. VANDER STOUW, I. NAZNITSKY, and JAMES E. RUSH
Chemical Abstracts Service, The Ohio State University, Columbus, Ohio 43210

Received June 14, 1967

Simultaneously with its development of a computer-based Chemical Compound Registry System, Chemical Abstracts Service is devising procedures for automatically converting systematic names of organic compounds into atom-bond connection tables which can be manipulated by computer. A study of systematic *Chemical Abstracts* (CA) index names has resulted in a dictionary of word roots used in the names and in step-by-step procedures for converting names to connection tables. Statistical studies of nomenclature in CA indexes show that these procedures are applicable not only to current nomenclature, but also generally to names in past indexes. Procedures have been written which are applicable to the majority of names of carbon compounds, and the preparation of computer programs is now under way.

NOMENCLATURE TRANSLATION AND ITS ROLE AT CAS

In 60 years, some 3.5 to 4 million chemical compounds associated with 15 to 20 million references have been reported in the world's scientific and technical literature and have been indexed in *Chemical Abstracts* by their systematic names. Each year, nearly 10% of this number—about 350,000 compounds—are processed for *CA* indexes; this figure includes both compounds that appear in the literature for the first time and old compounds that are reappearing in the literature. Identification of these compounds depends on a knowledge of their molecular structure, and the ability to describe molecular structure continues to be the basic factor assuring the value of the literature to today's practicing scientist.

In order to process this large and growing collection of data, Chemical Abstracts Service (CAS) established in 1965 a Chemical Compound Registry System. This computer-based system, which depends upon molecular structure as the means of identifying a compound, includes a computer record of the structure, names, molecular formula, and bibliographic citations for each compound registered. In less than two years of operation nearly 600,000 different chemical compounds have been identified and recorded in this system. Almost all of these compounds have been input *via* structural diagrams which were hand-drawn by professionals and translated into machine language through clerical effort *(1)*.

One requirement that must be met before the full benefit of this Registry System can be derived is that of including in the record information on all chemical compounds that have been reported in the literature. Only with a complete record, for example, will it be possible to determine with certainty that a given compound is or is not new to chemistry. Thus, the tremendous body of compound-oriented information that appears in pre-1965 indexes to *CA* constitutes an important collection for future registration. Nevertheless, this registration task will not be feasible—in terms of time, manpower, or dollars—if it is to require the preparation of a structural diagram for each compound and the subsequent conversion of that structural diagram to machine language for registration. Fortunately, past indexes to *CA* have indexed compounds through a systematic nomenclature that is based directly on the diagrammatic representation of structure which

*Presented in part before the Division of Chemical Literature, Symposium on Progress Report on Modern Methods, 153rd National Meeting of the American Chemical Society, Miami Beach, Fla., April 1967.

is universally used throughout chemistry. The direct relationship between structural diagrams and systematic chemical nomenclature provides the link that permits direct computer translation from nomenclature to connection tables for use in automatic registration. This task—the automatic translation of a systematic compound name into the corresponding computer-manipulable structural representation—constitutes what we call "Nomenclature Translation."

The possibility of nomenclature translation has previously been discussed by Dyson (*2*), Garfield (*3*), Opler (*4*), and several Russian scientists (*5*). However, no widely applicable set of rules or programs for conversion of names of organic compounds has yet been reported.

We here report the development of translation procedures applicable to a wide range of systematic nomenclature. These procedures, which are still being developed to encompass a broader range of compounds, will allow by the end of 1967 the handling of an estimated 80 to 85% of the names that appear in current and past *CA* Formula Indexes.

CRITERIA FOR AUTOMATIC NOMENCLATURE TRANSLATION

Automatic nomenclature translation procedures must meet several criteria if they are to be economical and practical for a large number of compounds.

First, and most obvious, translation procedures must be computer applicable—that is, they must constitute a set of logical rules applied unvaryingly and without requiring human judgment. Second, the procedures must be positive in their application so as to keep the error level to a minimum. This means that a name input to the translation program should either be translated correctly or rejected as untranslatable. Third, the procedures must be broadly applicable to a wide range of chemical names, including names which have not yet been written, but which will be in the future according to the rules of systematic nomenclature. Fourth, the procedures should require minimum intervention by professional and clerical personnel. This implies that the programs should be able to accept names keyboarded directly from *CA* indexes without requiring special coding, formatting, or abbreviation. Finally, an implied requirement in the large-scale translation of many names is that names be automatically edited upon entering the system to assure the accuracy of the input.

Our nomenclature translation procedures have been developed to meet these criteria. These procedures are based on the results of systematic analysis of extensive samples of names from the Formula Indexes to *CA*. These name samples, which together encompass more than 40,000 names, have been used in developing the procedures, testing them, gauging their range of applicability, and determining promising fields of chemical nomenclature for extending that range.

Any procedure for handling chemical nomenclature must be capable of accepting wide variety in the length and character make-up of names. The character set of which *CA* index nomenclature is made up contains 377 different characters; these include Roman, Greek, numeric, and punctuation characters, as well as superscripts, italics, small capitals, and so forth. Preferred index names can be as short as three letters or as long as several hundred, with no theoretical upper limit.

AN EXAMPLE OF NOMENCLATURE TRANSLATION

CAS nomenclature translation techniques work as follows:

Analysis proceeds as a name is read through, character by character, left to right. Locants and multipliers are identified and stored until needed; the carbon skeletons of the parent compound and of the radicals attached to it are identified through a small dictionary of basic word roots; these skeletons are modified and augmented through a series of extensive instructions based upon permissible combinations of prefixes, suffixes, and word roots. Subassemblies and assemblies of the developing structure are fabricated and joined with the aid of punctuation and numerical information in the name; temporary ambiguity is resolved by context or by reference to a special name list; and finally, a completed structure is read into the machine record for further processing. If at any point in the analysis, a name is found to be inconsistent, incomplete, or ambiguous, then that name will be rejected and reviewed by a chemist.

To illustrate this procedure the actual conversion of a name to a structural record will be described step by step. The example used here is a fairly simple one, so its translation proceeds in a relatively straight-forward manner without making use of many of the checks and special routines built into the translation procedures. In addition, some of the minor details and the repetitive processing steps have been simplified. Nevertheless, the example will serve to illustrate the basic operation of the CAS nomenclature translation procedures.

Figure 1(a) shows the name to be translated: 1-piperazinecarboxylic acid, 4-(3,4,5-trimethoxybenzoyl)-, ethyl ester. Like all *CA* systematic names, this name contains a section in which the parent compound is named. The computer will recognize this section as including all of the characters that precede the first comma followed by a blank. Following this portion of the name are other sections of the name that describe radicals and modifications; these sections too, are separated by commas followed by blanks.

The first step is a check for the presence of locants; such a check is made at any point where a locant can occur. This check identifies not only numeric locants, but also Greek letters and italicized letters when these serve as locants. In this name, the number 1 is found and is stored for later use. Next, there are checks as to whether the first two letters are "bi" or "di." Neither is found. Now the first three letters, "pip," are compared to the dictionary of word roots. The number of word roots in this dictionary is low; the total should not exceed 500 to 600. This list is small enough to keep in the computer memory during processing; thus, the translation programs can run at internal central processing-unit speeds; dictionary look-up time will not constitute a limiting factor.

When a word root is looked up in the dictionary, four possibilities arise. First, there may be no word roots on file which begin with the three letters. If so, translation will stop and the name will be put out for review. Second, there may be a three-letter word root on file which matches the three letters. In this case, the word root is retrieved, together with any special instructions as to subsequent processing steps. In the other two possibilities, the dictionary contains one or more longer word roots beginning with the three letters. In the name chosen as an example, the fourth possibility is true. The dictionary contains four word roots that begin with "pip": "pipecol," "piperon," "piperid," and "piperazin." Now the name is compared letter by letter to the four word roots until one is completely matched. If none is matched, the name is rejected.

Here the root "piperazin" is identified and the corresponding structure is set up with the conventional numbering. At this point in translation the locant 1 is stored and the parent ring structure set up, as in Figure 1(b). The identification and retrieval of this parent structure result in an instruction as to which routine is to be used in analyzing the ensuing suffixes. Special routines often have to be followed in order to resolve unusual or potentially ambiguous situations. In this case, the routine is one of wide applicability, since there are no unusual aspects to the usage of the word root "piperazin" requiring special treatment. This general routine proceeds first with a check for the letter "e," which is found in this case. Since "e" serves merely for pronunciation, its identification does not lead to any alteration in structure. The next check is for the multiplier "di," which is not found. Now, the three letters "car," which follow the letter "e," are compared to a small number of letter sequences permitted to occur at this point. Among the other possibilities are the multiplier "tri" and the letters "sul," leading into the sulfonic acids and their derivatives. If none of the permissible letter sequences occur, translation will stop and the name will be put out for review. Once "car" has been identified, the remainder of the parent compound name is compared to the letter sequences that can possibly follow it. Other possible functional group names besides "carboxylic acid" include "carboxaldehyde," "carbonyl chloride," and so on.

When the suffix "carboxylic acid" has been completely identified, an instruction is given to place a carboxyl group in the structure. Placement is done by a standard routine which uses the locants and/or multipliers that are held to control placement of the new groups. For this name, the locant 1 is held, and placement of the carboxyl group gives the structure shown in Figure 1(c). The routine used for placing such groups involves several editing checks—for example, a name without sufficient locants, or one in which the number of locants and a multiplying term did not agree, would ordinarily be rejected at this point. Similarly, if placement of the group created a pentavalent carbon atom or a tetravalent nitrogen atom, the name would be rejected unless some other feature, such as a positive charge on the nitrogen, allowed the situation.

Analysis now proceeds with the second segment of the name, in which the radicals are described. First the locant "4" is identified and held. Then an open curve is found. This is interpreted to mean that the terms which appear between that point and the corresponding closing curve constitute the name of a complex branch structure. Each of the structural fragments that corresponds to a term between the curves is held separately until the closing curve is reached. Then the last fragment named within the curves is considered to be the main part of the complex branch; the other fragments are all attached to it. The steps involved in converting the term in curves are shown in Figure 1(d). First the locants 3, 4, and 5 are identified and held. Since there are three locants, a check is made whether the next three letters are the corresponding multiplier "tri," which is found and held. Then the terms "methoxy" and "benzoyl" are treated through dictionary look-up of their basic word roots "meth" and "benz" plus an analysis of the suffixes to modify the basic structures. Once the closing curve is reached, the branch is assembled by treating the last radical to be named, the "benzoyl" radical, as a parent structure and the earlier radicals as branches of it. With the use of the multiplier "tri" and the three locants, the branch is assembled as shown.

After the closing curve, there is found a dash, a comma, and a blank, indicating the end of the chain of radical names. Now the locant "4," which preceded the parenthetical term, is used to attach the branch to the parent structure, giving the assembled structure shown in Figure 1(e).

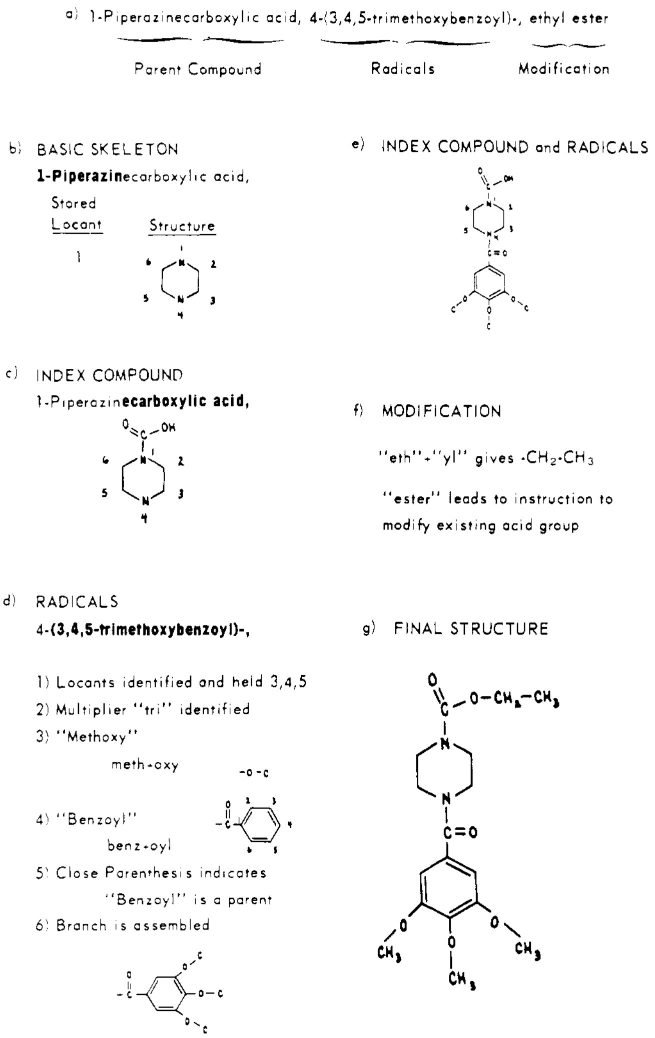

a) 1-Piperazinecarboxylic acid, 4-(3,4,5-trimethoxybenzoyl)-, ethyl ester
Parent Compound
Radicals
Modification
b) BASIC SKELETON
1-Piperazinecarboxylic acid,
Stored Locant
Structure
1
c) INDEX COMPOUND
1-Piperazinecarboxylic acid,
d) RADICALS
4-(3,4,5-trimethoxybenzoyl)-,
1) Locants identified and held 3,4,5
2) Multiplier "tri" identified
3) "Methoxy"
meth+oxy
-o-c
4) "Benzoyl"
benz+oyl
5) Close Parenthesis indicates "Benzoyl" is a parent
6) Branch is assembled
e) INDEX COMPOUND and RADICALS
f) MODIFICATION
"eth"+"yl" gives ·CH₂·CH₃
"ester" leads to instruction to modify existing acid group
g) FINAL STRUCTURE

Figure 1. Example of nomenclature translation.

Analysis of the term "ethyl ester" proceeds in a fashion similar to the analysis of the radical names. The three letters "eth" are identified through dictionary look-up and the appropriate structure is provided. Identification of the term "ester" leads to an instruction to modify the existing acid group by replacing a hydrogen by an ethyl radical. In order to facilitate this type of operation, the translation algorithm keeps a record of the nature and location of functional groups likely to be modified, such as carboxylic acid groups, carbonyl groups, etc. Here it is found that a carboxylic acid group is attached to atom number 1. Replacement of the hydrogen by an ethyl radical gives the final structure shown in Figure 1(g).

COMPLICATING FACTORS

While the example just discussed is a fairly straightforward *CA* name, the translation procedures that have been developed can be applied to a wide variety of more complex situations. Consideration of some other names which can be handled will help to give a qualitative picture of the present scope of the translation procedures.

One complication that can arise in name translation involves word roots that deviate from the systematic pattern followed by other members of the same family. Such deviations can usually be attributed to a common usage retained by *CA*, often because it aids in pronunciation or helps to avoid ambiguity. Some commonly occurring examples of such deviations are shown in Figure 2. In the first example, *CA* uses "pyridone" rather than "pyridinone" because of traditional usage. In the second case, the extra "e" is retained in "thiophene-3-ol" to avoid possible ambiguity with "thiophenol." The translation procedures allow for a number of such common usages.

CA NAME	"SYSTEMATIC" NAME
Pyridone	Pyridinone
Thiophene-3-ol	Thiophen-3-ol 3-Thiophenol
Methacrylic acid	Acrylic acid, α-methyl-

Figure 2. Complicating factors in deviations from systematic usage.

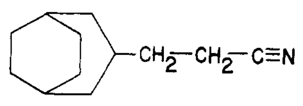
It has also been necessary to allow for word roots that have different meanings in different settings. Two examples of such word roots are:

Cyclo**hex**anol, 2,2,4,4,6,6-**hex**achloro-

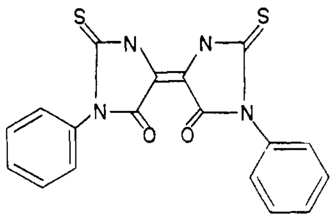
Hydantoin, 3-[4-(dimethylamino)-3,5-dinitrophenyl]-5-[2-(methyl**thio**ethyl]-2-**thio**-

In the first name, the first occurrence of "hex" refers to six carbon atoms; the second "hex" acts as a multiplier and would be so identified. In the second name, the first "thio" represents a sulfur atom which connects two carbon atoms *via* single bonds; the second "thio" refers to a sulfur atom which replaces a previously described oxygen atom. A special routine resolves this potential ambiguity in terms of the settings in which "thio" is found.

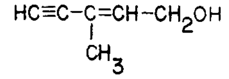
Bicyclo[3.2.2]nonane-3-propionitrile



[$\Delta^{4,4'}$-Biimidazolidine]-2,2',5,5'-tetrone, 1,1'-diphenyl-2,2'-dithio-



Structure



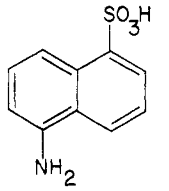


Figure 3. Complicating factors in "bi" nomenclature.

In Figure 3, two names are given which involve the prefix "bi." In the first case, the set of numbers in brackets, "[3.2.2]" is used by the procedures to construct the basic carbon skeleton; the number of carbon atoms which results is then checked against the number of carbon atoms that correspond to the following word root, "non." A discrepancy would cause the name to be rejected.

In the other "bi" name, the locant pair "$\Delta^{4,4'}$" and the prefix "bi" are held until the closing is reached; then the imidazolidine structure is doubled, the second structure is numbered using primed numbers, and the two structures are attached *via* a double bond between the 4 and 4' positions.

"1-(2-Chloro-4-phenylbenzyl)pyridinium bromide" is an uninverted name as used in *CA* Formula Indexes to describe organic ionic compounds. The treatment of such a name is similar to the handling of complex branch names that has already been illustrated: The individual radical structures are set up and are held separately until the parent structure has been identified; then all of the branches are attached to the parent structure.

The name "2*H*-Cyclopenta[5,6]naphth[1,2-*d*]oxepin" is based on the name of a fused ring system, where the ring system is named according to what is sometimes called the Patterson system. In this method of naming ring systems, the position of fusion of two or more simple ring systems is designated by an alphanumeric term in brackets, such as "[5,6]" or "[1,2-*d*]." The direct conversion of such fused ring names to connection tables is theoretically possible, and approaches to this problem have indeed been suggested by Dyson (*2*) and by Tsukerman (*5d*). However, the required routines would be inordinately complex in relation to the proportion of names that actually use this convention. We have therefore chosen to handle these systems by the use of a list of about 1000 of the most common names of fused ring systems and their corresponding connection tables. In a name such as that shown here, the presence of an open bracket immediately after the term "cyclopenta" will lead to an instruction to go to the list of fused systems. In this case, there would be found on the list the system "2*H*-cyclopenta[5,6]naphth[1,2-*d*]oxepin." The connection table for the ring system will be brought out, and the remainder of the name analyzed by the techniques already described to give the final structure.

NOMENCLATURE TRANSLATION AS A NAME EDITING TOOL

Nomenclature translation techniques also serve as diagnostic editing devices for chemical names. The nomenclature translation techniques offer powerful computer-based support to the professional scientist who handles chemical nomenclature in large volume. For example, as Garfield has shown (*3*), it is feasible to determine the molecular formula from the name by computer. This resulting formula can then be compared with the value calculated and entered by a chemist—a good test for self-consistency and for consistency between the name and the structure. The computer can also be programmed to identify and correct certain kinds of typographical errors and to ensure consistency in the use of capitals and italics. CAS has developed a list of about 50 prefixes such as "arabino" and "meso" that are automatically italicized by the computer whether or not they have been italicized on input. Similarly, the computer will identify the first letter in the main portion of the name and capitalize it so that two names are not counted as different when in fact they differ only by the use of capitalization. Other programs will expand line formulas and element symbols such as HCl for hydrochloride and Na for sodium.

Finally, the nomenclature translation procedures will not process an incomplete name, an inconsistent name, or an ambiguous name. When a name is fully processed and converted to a structural record, a molecular formula can be calculated from that record and compared with the chemist-calculated molecular formula which has also been input; any discrepancy will cause the name to be rejected for review. The rejection of these names by the algorithm, together with diagnostic messages indicating why and where the name was rejected, constitute a significant aid to the editing and correction of such names. In all of these editing checks, there is a very close parallel to the diagnostic routines for accuracy that are part of the structure registry operations and that have been described by Leiter and Morgan (*6*).

APPLICABILITY OF THE TRANSLATION PROCEDURES

To obtain a quantitative picture of the extent to which these procedures can be applied to *CA* Index Nomenclature, the procedures have been tested on two statistically designed samples of *CA* names. These included one sample of about 4000 names of carbon compounds from the Formula Index to Volume 61 and another of the same size chosen from Volume 51. Procedures are at present applicable to 62% of the names from Volume 61 and 58% from Volume 51. In terms of the elemental make-up of the compounds covered, the procedures are at present generally applicable to names of compounds of carbon with any combination of these elements: Br, Cl, F, H, I, N, O, P, S, Si.

Three important types of nomenclature remain to be studied this year. These are the names of amino acids, carbohydrates, and of compounds containing boron or metals. Once these studies have been completed, the translation procedures should be applicable to 80 to 85% of names of carbon-containing compounds. This probably represents a practical upper limit; beyond this point there are a wide variety of special problems for which the law of diminishing returns sets in.

The examples of names used in this paper have all been *CA* preferred Index names, but the translation procedures are not limited to such names. They can be applied to any unambiguous name formed from the word roots in the dictionary according to the rules of systematic nomenclature that are recognized by the routines. In Figure 4 are shown two compounds with a set of names for each. The systematic names shown can be handled by these procedures, both in their inverted and uninverted forms, but the nonsystematic or trivial names will not be translated. One could, of course, write more names for each of these compounds; those recognized by the routines would be translated, while those not recognized would be rejected.

Names Translated	
2-Penten-4-yn-1-ol, 3-methyl-	1-Naphthalenesulfonic acid, 5-amino-
2-Penten-4-yne, 3-methyl-1-hydroxy-	1-Naphthylamine, 5-sulfo-
Pent-3-en-1-yne, 5-hydroxy-3-methyl-	Naphthalene, 1-amino-5-sulfo-
Names Rejected	
1-Pentol	1,5-Amino Acid
	Laurent's Acid

Figure 4. Translation of more than one name for a structure.

ACKNOWLEDGMENT

We gratefully acknowledge the support provided to this work by the National Science Foundation, Department of Defense, and the National Institutes of Health. We also acknowledge the impetus given the work through early discussions with Dr. Michael F. Lynch.

LITERATURE CITED

(1) Leiter, D. P., Jr., Morgan, H. L., and Stobaugh, R. E., J. CHEM. DOC. **5**, 238 (1965).

(2) Dyson, G. M., *Inform. Storage Retrieval* **2**, 59 (1964).

(3) *a.* Garfield, E., J. CHEM. DOC. **2**, 177 (1962).
b. Garfield, E., *Nature* **192**, 192 (1961).

(4) Opler, A., *Amer. Doc.* **10**, 59 (1958).

(5) *a.* Tsukerman, A. M., and Terentiev, A. P., *Proc. Intern. Conf. on Standards for a Common Language for Machine Searching and Translation* **1**, 493 (Interscience Press, 1960).
b. Seifer, A. L., and Shtein, V. S., *Nauch.-Tekh. Inform., Vses. Inst. Nauch. Tekh. Inform.* **1960**, No. 1, 172.
c. Stetsyura, G. G., and Tsukerman, A. M., *ibid.*, **1962**, No. 3, 17–19.
d. Tsukerman, A. M., *ibid.*, **1965**, No. 4, 23–30.
e. Seifer, A. L., *Kibern. Dok., Akad. Nauk S.S.S.R.* **1966**, 101–37.

(6) Leiter, D. P., Jr., and Morgan, H. L., J. CHEM. DOC. **6**, 226 (1966).