Table I. Enumeration of Polyhexes with up to $h = 16$

		planar polyhexes with holes		CPU time			
h	planar polyhexes without holes	single hole	multiple holes	days	h	min	s
1	1						0.44
2	1						0.44
3	3						0.44
4	7						0.44
5	22						0.60
6	81						1.26
7	331						4.67
8	1 435	1					20.87
9	6 505	5				1	40.13
10	30 086	43				8	12.90
11	141 229	283				40	52.17
12	669 584	1 954			3	24	1.80
13	3 198 256	12 363	1		17	4	32.50
14	15 367 577	76 283	11	3	14	4	24.76
15	74 207 910	453 946	149	18	3	13	34.70
16	359 863 778	2 641 506	1618	91	7	24	33.69

tween the terms polyhex and benzenoid.

Our algorithm for enumeration of polyhexes is based on the DAST (the dualist angle-restricted spanning tree) code.[13] This code is founded on the graph-theoretical concept of the weighted spanning tree of dualist.[14] A computer program for enumerating polyhex hydrocarbons using our algorithm is detailed elsewhere.[8] Counts of planar polyhex hydrocarbons with and without holes with up to $h = 16$ are given in Table I. In the table we also give CPU time needed to complete computation for each h. For example, the required CPU time for the enumeration of polyhexes with 16 hexagons was 91 days, 7 h, 24 min, and 33.69 s.

Computations have been carried out on the Siemens PCD3D (20 MHz, 386-AT). We deliberately did not use our supercomputer to show that these types of combinatorial computations can be completed on the personal computer, if one is willing to spend some time to design carefully an efficient computational algorithm and if one can spare a PC for continuous computations over a long period of time.

ACKNOWLEDGMENT

This work was supported in part by the German-Yugoslav Scientific Cooperation Program. Financial support from the Internationales Büro, Kernforschungsanlage—Jülich, and from the Croatian Science Fund is gratefully acknowledged. We are thankful to reviewers for their critical but helpful comments.

REFERENCES AND NOTES

(1) Tošić, R.; Kovačević, M. Generating and Counting Unbranched Catacondensed Benzenoids. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 29-31.

(2) Brunvoll, J.; Cyvin, B. N.; Cyvin, S. Enumeration and Classification of Coronoid Hydrocarbons. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 14-21.

(3) Brunvoll, J.; Cyvin, B. N.; Cyvin, S. Enumeration and Classification of Benzenoid Hydrocarbons. 2. Symmetry and Regular Benzenoids. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 171-177.

(4) Cyvin, S.; Brunvoll, J.; Cyvin, B. N. Distribution of *K*, the Number of Kekulê Structures, in Benzenoid Hydrocarbons: Normal Benzenoids with *K* to 110. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 74-90.

(5) Ciosłowski, J. Computer Enumeration of Polyhexes Using the Compact Naming Approach. *J. Comput. Chem.* **1987**, *8*, 906-915.

(6) Harary, F.; Palmer, E. M. *Graphical Enumeration*; Academic: New York, 1973.

(7) Balasubramanian, K.; Kauffman, J. J.; Koski, W. S.; Balaban, A. T. Graph Theoretical Characterization and Computer Generation of Certain Carcinogenic Benzenoid Hydrocarbons and Identification of Bay Regions. *J. Comput. Chem.* **1980**, *1*, 149-157.

(8) Müller, W. R.; Szymanski, K.; Knop, J. V.; Nikolić, S.; Trinajstić, N. On the Enumeration and Generation of Polyhex Hydrocarbons. *J. Comput. Chem.* (in press).

(9) Trinajstić, N. *Chemical Graph Theory*; CRC: Boca Raton, FL, 1983; Vol. 1, Chapter 3.

(10) Trinajstić, N. On the Classification of Polyhex Hydrocarbons. *J. Math. Chem.* (in press).

(11) Harary, F. *Graph Theory*; Addison-Wesley: Reading, MA, 1971; second printing.

(12) Knop, J. V.; Müller, W. R.; Szymanski, K.; Trinajstić, N. On the Enumeration of 2-Factors of Polyhexes. *J. Comput. Chem.* **1986**, *7*, 547-564.

(13) Knop, J. V.; Müller, W. R.; Szymanski, K.; Nikolić, S.; Trinajstić, N. Computer-oriented Molecular Codes. In *Computational Chemical Graph Theory*; Rouvray, D. H., Ed.; Nova: New York, 1990; pp 9-32.

(14) Nikolić, S.; Trinajstić, N.; Knop, J. V.; Müller, W. R.; Szymanski, K. On the Concept of the Weighted Spanning Tree of Dualist. *J. Math. Chem.* (in press).

Molecular Topological Index

WOLFGANG R. MÜLLER, KLAUS SZYMANSKI, JAN V. KNOP, and NENAD TRINAJSTIĆ*,[‡]

Computer Centre, The Heinrich Heine University, 4000 Düsseldorf, Federal Republic of Germany

Received November 13, 1989

The molecular topological index (MTI) has been systematically tested for counterexamples (two or more nonisomorphic structures with the same MTI number). The analysis was carried out for alkane trees with up to 16 atoms. The search for counterexamples was positive: The first pair of alkane trees with identical MTI numbers was found in the octane family. The more disturbing finding was that two nonisomorphic alkane trees of different sizes may also possess the same MTI value. An attempt to redefine the MTI in terms of only the distance matrix and the valency matrix was abortive.

Schultz[1] has recently introduced in this journal a novel topological (or, more correctly, graph-theoretical) descriptor for characterization of alkanes by an integer. This descriptor was named by its originator the molecular topological index (MTI). The MTI appears to be an attractive graph-theoretical descriptor that is easy to compute and has structural significance. The important questions, Are the MTI numbers unique?, and, if the answer is positive, What are the smallest graphs for which MTI is not unique?, however, were not considered by Schultz. He only stated that the MTI is a highly discriminative descriptor. In this paper we will try to answer these questions.

The MTI is based on the adjacency ($N \times N$) matrix A, the distance ($N \times N$) matrix D, and the valency ($1 \times N$) matrix v of an alkane. The sum of the elements e_i ($i = 1, 2, ..., N$)

[‡] Permanent address: The Rugjer Bošković Institute, P. O. Box 1016, 41001 Zagreb, Croatia, Yugoslavia.

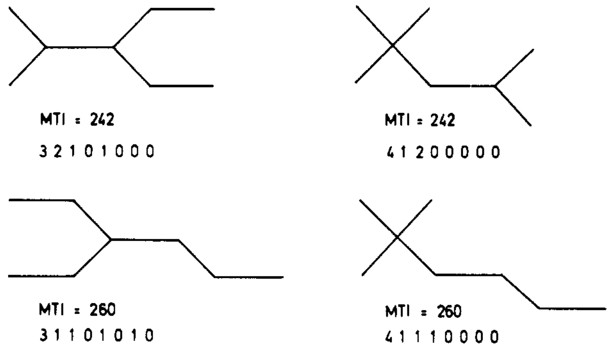

Figure 1. Two pairs of octane trees with degenerate MTI values. Beneath each tree its N-tuple code is also given.

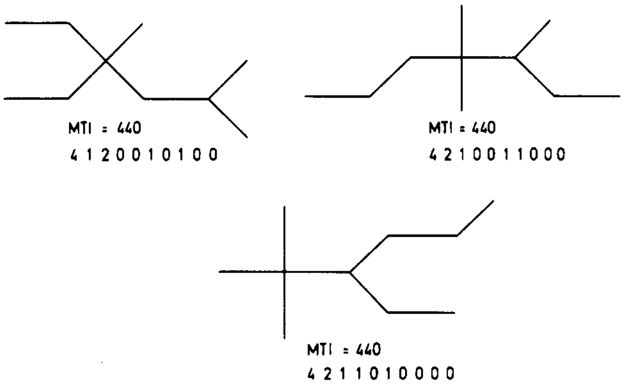

Figure 2. Triplet of decane trees with the same MTI value. Beneath each tree its N-tuple code is also given.

of the product of matrices $v(A + D)$ gives the molecular topological index:

$$\mathrm{MTI} = \sum_{i=1}^{N} e_i$$

For example, the product $v(A + D)$ for 2-methylpropane is given by

$$v(A + D) = [10\ 6\ 10\ 10]$$

The summation of the elements in the row matrix above produces the MTI of 2-methylpropane:

$$\mathrm{MTI} = 10 + 6 + 10 + 10 = 36$$

The matrix $A + D$ which represents the sum of the adjacency matrix and the distance matrix may be directly constructed from the labeled graph (tree) of an alkane. One has simply to input into the distance matrix of a tree digit 2 in all places where the distance of unity would be otherwise placed.

Since we are in the position to generate routinely all alkanes up to a given number of carbon atoms[2-4] and have ready a computer program for construction of their distance matrices,[5] we computed the MTI numbers for all alkane trees with up to 16 atoms (altogether 18 030 alkanes) in our search for nonisomorphic trees with identical MTI values. Our target was achieved very early in the computation. Two pairs of alkane trees (corresponding to 3-ethyl-2-methylpentane and 2,2,3-trimethylpentane and to 3-ethylhexane and 2,2-dimethylhexane) with the same MTI numbers appeared in the octane family (18 members). These alkane trees, which represent the smallest pairs of trees with degenerate MTI values, are given in Figure 1. All alkane trees in Figure 1 and in other figures that follow will be also identified by their N-tuple codes[2] because we used these codes to generate trees.

We found six pairs of trees in the nonane family (35 members) and 15 pairs of trees in the decane family (75 members) with the same MTI values. They are listed in Table I. Each tree in the table is identified by its N-tuple code.

Table I. Duplicate MTI Values for Nonane and Decane Trees[a]

N-tuple code	MTI
(a) Nonanes	
4 2 1 0 0 1 0 0 0	318
4 2 1 0 1 0 0 0 0	318
3 2 2 1 0 0 0 0 0	332
4 1 1 0 1 0 1 0 0	332
4 1 2 0 0 1 0 0 0	334
4 2 1 1 0 0 0 0 0	334
3 2 1 2 0 0 0 0 0	348
4 1 1 0 1 1 0 0 0	348
3 1 2 1 0 1 0 0 0	354
3 2 1 1 0 0 1 0 0	354
3 1 2 1 1 0 0 0 0	370
3 2 1 1 1 0 0 0 0	370
(b) Decanes	
4 2 0 0 2 0 0 1 0 0	402
4 2 3 0 0 0 0 0 0 0	402
4 2 1 0 0 2 0 0 0 0	414
4 2 2 0 0 1 0 0 0 0	414
3 2 2 0 0 2 0 0 0 0	420
4 2 1 0 0 1 0 1 0 0	420
4 1 1 0 1 0 1 0 1 0	434
4 2 0 0 1 2 0 0 0 0	434
3 2 2 0 0 1 1 0 0 0	446
4 2 1 2 0 0 0 0 0 0	446
3 2 2 1 0 0 0 1 0 0	450
4 1 2 2 0 0 0 0 0 0	450
4 1 2 1 0 0 1 0 0 0	456
4 1 2 1 0 1 0 0 0 0	456
4 1 2 0 0 1 1 0 0 0	460
4 2 0 0 1 1 1 0 0 0	460
3 2 1 1 0 1 0 1 0 0	464
4 1 1 3 0 0 0 0 0 0	464
3 2 1 1 0 1 1 0 0 0	472
4 2 1 1 1 0 0 0 0 0	472
4 1 1 2 0 0 1 0 0 0	476
4 1 2 1 1 0 0 0 0 0	476
3 1 2 1 1 0 1 0 0 0	484
3 2 1 1 1 0 1 0 0 0	484
3 2 1 1 0 0 1 1 0 0	488
4 1 1 2 1 0 0 0 0 0	488
3 1 1 2 1 0 1 0 0 0	500
3 1 2 1 1 0 0 1 0 0	500
3 1 1 2 1 1 0 0 0 0	520
3 2 1 1 1 1.0 0 0 0	520

[a] Each tree is identified by its N-tuple code.

The first triplet of alkane trees (corresponding to 3-ethyl-3,5-dimethylhexane, 2,3,3-trimethylheptane, and 3-ethyl-2,2-dimethylhexane) with degenerate MTI values appeared in the decane family. These trees are given in Figure 2.

The results above indicate that MTI is an index of low discriminatory power. However, this alone would not be considered a weakness of the MTI if it could be established as a useful descriptor in various applications. The truism is that the validity of a given molecular descriptor (graph-theoretical index) depends on its usefulness. This is best exemplified in the case of the connectivity index (χ).[6] Since Schultz compared the MTI with χ in his work, we point out that χ is also an index of low discriminatory power. In the octane

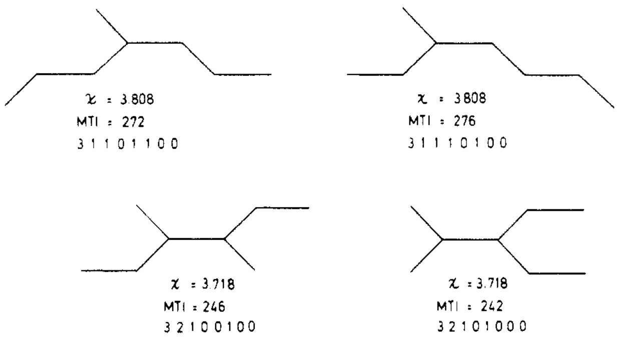

Figure 3. First two pairs of alkane trees with identical values of connectivity indices. Beneath each tree its MTI number and N-tuple code are also given.

Table II. Minimum and Maximum MTI Values for Alkane Trees with up to 16 Vertices

no. of vertices	no. of alkane trees	MTI no. minimum	MTI no. maximum
1	1	0	0
2	1	4	4
3	1	16	16
4	2	36	38
5	3	64	74
6	5	106	128
7	9	156	204
8	18	214	306
9	35	298	438
10	75	390	604
11	159	490	808
12	355	616	1054
13	802	750	1346
14	1858	892	1688
15	4347	1060	2084
16	10359	1236	2538

family appear two pairs of alkane trees (corresponding to 3-methylheptane and 4-methylheptane and to 2-ethyl-3-methylhexane and 3-ethyl-2-methylpentane) with identical connectivity indices. These structures are given in Figure 3.

The above finding did not harm the connectivity index. χ, in spite of being a low discriminatory index, has been used in various quantitative structure–property relationship (QSPR) and quantitative structure–activity relationship (QSAR) studies,[7,8] more than any other of 120 or so graph-theoretical indices that have been so far proposed in the literature,[9] because of its many attractive features. For example, it could be easily extended to cover the heterosystems.[7,8] Incidentally, the MTI may also be simply extended to heterosystems because the procedures for setting up the adjacency matrices and the distance matrices of heterosystems are already available.[10,11]

The following finding, however, is not exhibited either by the connectivity index or by a number of other graph-theoretical indices. We have computed the minimum and maximum MTI numbers [$(\mathrm{MTI})_{\min}$ and $(\mathrm{MTI})_{\max}$] for families of alkane trees with up to 16 vertices. These results are reported in Table II. From Table II we learn that $(\mathrm{MTI})_{\min}$ and $(\mathrm{MTI})_{\max}$ overlap between the alkane families with eight vertices and higher. This result indicates the possibility that two nonisomorphic alkane trees of different sizes may possess the same MTI number. Indeed, as we predicted we found three such cases connecting nonanes and decanes (3-methyloctane and 2-ethyl-2,3,3-trimethylpentane, 2-methyloctane and 2,2,3,3-tetramethylhexane, and nonane and 3-ethyl-2,4-dimethylhexane). These structures are given in Figure 4. This finding disproves the statement that the MTI is essentially monotonic.[1] The above analyses indicate that the MTI exhibits two kinds of nonuniqueness: (i) Nonisomorphic trees of the same size may have the same value of MTI, and (ii) nonisomorphic trees of different sizes may also have the same MTI numbers.

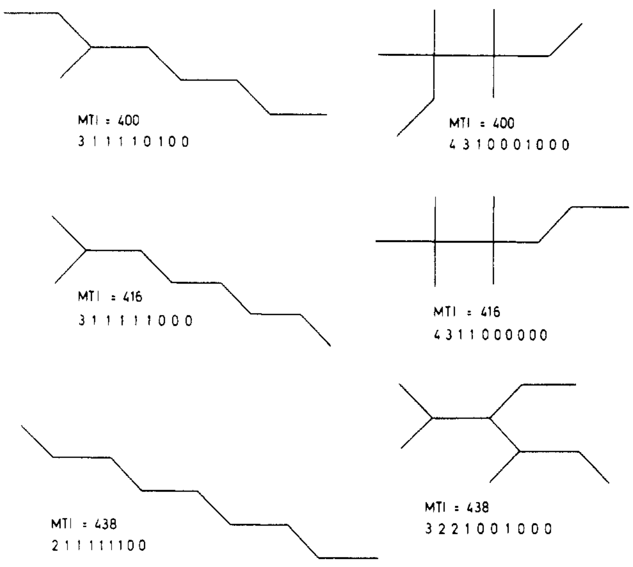

Figure 4. Three pairs of alkane trees of different sizes with the same MTI number. Beneath each tree its N-tuple code is also given.

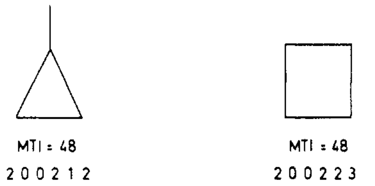

Figure 5. Two cyclic structures possessing the same MTI number. Beneath each structure its compact code is also given.

Although Schultz stated[1] that the MTI is an index for characterization of alkanes, we applied it to cycloalkanes and their derivatives. Almost immediately in our analysis we encountered a pair of cyclic structures (corresponding to methylcyclopropane and cyclobutane) with the same MTI number. These two structures represent the smallest pair of graphs with identical MTI values. They are depicted in Figure 5. Each structure is also identified by its compact code.[12,13] Compact codes are structural codes resulting from an extension of the N-tuple code to cyclic systems.

We also attempted to redefine the MTI. The comparison between the matrices $A + D$ and D indicates that the distance matrix D is a richer mathematical structure (possesses higher information content)[14] than the matrix $A + D$ because in the latter matrix there is not any more distinction made between the graph-theoretical distances of length 1 and length 2. Therefore, we redefined the MTI to be a sum of elements e_i ($i = 1, 2, ..., N$) of the product of matrices v and D and denoted it (MTI)′. For example, the product vD for 2-methylpropane is given by

$$vD = [7\ 3\ 7\ 7]$$

and consequently the (MTI)′ of 2-methylpropane is 24.

We introduced the (MTI)′ with the hope that this descriptor will be more discriminative than the MTI, because the distance matrix D has higher information content than the sum matrix $A + D$. However, this conjecture was not supported by numerical analysis. In the heptane family two pairs of nonisomorphic trees (corresponding to 2,3-dimethylpentane and 2,2-dimethylpentane and to 3-ethylpentane and 2,4-dimethylpentane) are found with the same (MTI)′ numbers. These trees are given in Figure 6. We have also found 2 pairs of trees in the octane family, 4 pairs in the nonane family, and 12 pairs in the decane family with identical (MTI)′ values.

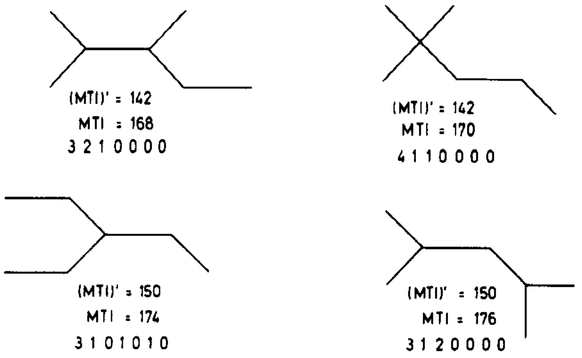

Figure 6. Two pairs of heptane trees with degenerate (MTI)′ values. Beneath each tree its MTI number and N-tuple code are also given.

Similarly, we found three triplets of trees in the nonane family and four triplets of trees in the decane family with identical (MTI)′ numbers. Finally, two quadruplets of trees in the nonane family and five quadruplets of trees in the decane family are found to possess the same (MTI)′ numbers. These results show that the (MTI)′ is an index of much lower discriminatory power than the MTI.

Finally, we considered the row matrices $v(A + D)$ as possible N-tuple descriptors. The results of the analysis, which was limited to 18030 alkane trees and several thousand cyclic structures, was that the row matrices are unique for the set of graphs investigated.

ACKNOWLEDGMENT

This work was supported in part by the German–Yugoslav Scientific Cooperation Program. Financial support from the Internationales Büro, Kernforschungsanlage—Jülich, and from the Croatian Science Fund is gratefully acknowledged. We thank Professor Harry P. Schultz (Coral Gables, FL) and the referees for their comments.

REFERENCES AND NOTES

(1) Schultz, H. P. Topological Organic Chemistry. 1. Graph Theory and Topological Indices of Alkanes. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 227–228.

(2) Knop, J. V.; Müller, W. R.; Jeričević, Ž.; Trinajstić, N. Computer Enumeration and Generation of Trees and Rooted Trees. *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 91–99.

(3) Trinajstić, N.; Jeričević, Ž.; Knop, J. V.; Müller, W. R.; Szymanski, K. Computer Generation of Isomeric Structures. *Pure Appl. Chem.* **1983**, *55*, 379–390.

(4) Knop, J. V.; Müller, W. R.; Szymanski, K.; Trinajstić, N. *Computer Generation of Certain Classes of Molecules*; SKTH/Kemija u industriji: Zagreb, 1985.

(5) Müller, W. R.; Szymanski, K.; Knop, J. V.; Trinajstić, N. An Algorithm for Construction of the Molecular Distance Matrix. *J. Comput. Chem.* **1987**, *8*, 170–173.

(6) Randić, M. On the Characterization of Molecular Branching. *J. Am. Chem. Soc.* **1975**, *97*, 6609–6615.

(7) Kier, L. B.; Hall, L. H. Molecular Connectivity in Chemistry and Drug Research; Academic: New York, 1976.

(8) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure–Activity Analysis*; Wiley: New York, 1986.

(9) Rouvray, D. H. The Limits of Applicability of Topological Indices. *J. Mol. Struct.* (*THEOCHEM*) **1989**, *185*, 187–201.

(10) Trinajstić, N. *Chemical Graph Theory*; CRC: Boca Raton, FL, 1983; Vol. 1, Chapter 4.

(11) Barysz, M.; Jashari, G.; Lall, R. S.; Srivastava, V. K.; Trinajstić, N. On the Distance Matrix of Molecules Containing Heteroatoms. In *Chemical Applications of Topology and Graph Theory*; King, R. B., Ed.; Elsevier: Amsterdam, 1983; pp 222–230.

(12) Randić, M. Compact Molecular Codes. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 136–148.

(13) Randić, M.; Nikolić, S.; Trinajstić, N. Compact Molecular Codes for Polycyclic Systems. *J. Mol. Struct.* (*THEOCHEM*) **1988**, *165*, 213–228.

(14) Bonchev, D.; Trinajstić, N. Information Theory, Distance Matrix and Molecular Branching. *J. Chem. Phys.* **1977**, *67*, 4517–4533.

Extraction of Chemical Reaction Information from Primary Journal Text

C. S. AI, P. E. BLOWER, JR.,* and R. H. LEDWITH

Chemical Abstracts Service, Columbus, Ohio 43210

Received November 28, 1989

This paper describes a series of programs that generate a summary of the preparative reactions reported in the experimental section of a paper in *The Journal of Organic Chemistry*. This summary identifies each participating substance along with its reaction role and quantity. It also records some procedural information such as the order of mixing the reactants, reagents, and solvents, and the duration and temperature of individual reaction steps. The work described here is potentially important as a means of automatically extracting reaction information from the ACS primary journal database and generating records for CASREACT with little intervention from CAS editorial staff.

INTRODUCTION

In earlier work with *The Journal of Organic Chemistry* (JOC), Zamora[1] showed that the techniques of computational linguistics can be used to extract facts about chemical reactions from the text of primary journals of the American Chemical Society (ACS). There is the potential to create useful, new databases from the existing ACS primary journal database with little or no extra editorial effort. The area of reactions was selected for our initial study for two reasons: (1) Analysis of the discriptions of synthetic preparations reported in the experimental section of a journal paper seems to be the right level of difficulty for early experiments using computational linguistics techniques for extracting information. Although the subject matter is restricted and the method of presentation is quite stylized and predictable, these descriptions still use natural language which exhibits considerable variation. (2) There is a potential for using the results to meet a real current need in building the file of reactions for CASREACT.

CASREACT is an STN[2] service that affords end-users access to chemical reactions reported in the current literature. To provide a database for this service, analysts in the Organic Chemistry department of CAS have been entering reaction data from more than 100 journals since October 1984. Locating and recording the necessary reaction information requires a very detailed level of analysis and is consequently labor intensive and time-consuming. Furthermore, the information recorded for a reaction only identifies the participating substances and the role of each. Except for product yield, all

0095-2338/90/1630-0163$02.50/0 © 1990 American Chemical Society