Essays of an Information Scientist, Vol:7, p.441-513, 1984

An Algorithm for Translating Chemical Names to Molecular Formulas

EUGENE GARFIELD

INSTITUTE FOR SCIENTIFIC INFORMATION
33 SOUTH 17th STREET • PHILADELPHIA 3, PENNSYLVANIA

Originally presented to the Faculty of the Graduate School of Arts and Sciences of the University of Pennsylvania in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy, 1961.

Library of Congress Catalog Card Number 61–17455

Institute for Scientific Information
33 South 17th Street • Philadelphia 3, Pennsylvania

PRINTED IN THE UNITED STATES OF AMERICA

PREFACE TO THE FIRST ISI EDITION

This varityped version of my doctoral dissertation has been prepared primarily to satisfy the many requests I received for copies of the original manuscript. With the exception of minor typographical changes and those noted below in the section on Transformations, the only other changes have been in the arrangement of the indexes, bibliography, etc. which had to conform to university conventions. However, in this edition the indexes, etc. have been placed at the end.

The original manuscript was typed primarily by my secretary, Mrs. Sylvia Shapiro. The varityping in this edition was done by Mrs. Joan M. Graham. Proofreading was performed by Mrs. Joan E. Shook and Mr. Walter Fiddler. Mr. Fiddler found errors of omission in the section on Transformations which have been corrected by the addition of footnotes. He also found many errors in the copying of chemical names and formulas in both the original and the final manuscript. This only strengthens my belief that an arduous intellectual task such as naming a chemical or calculating its formula is most consistently performed by a machine.

I also want to thank collectively, the many other persons who helped in the preparation of this work through suggestions and participation. The dissertation, as accepted by the Department of Linguistics of the University of Pennsylvania, went through several revisions before it was accepted. Many of these changes resulted from different interpretations of the *morpheme*, *allomorph*, etc. Linguistics is not yet so precise that one can prescribe a discovery procedure. Quite simply, this means that linguistic data can be interpreted in many useful ways. For the reader who is interested in pursuing the theoretical background of this statement further, I recommend Noam Chomsky's *Syntactic Structures* (Mouton & Co. 'S-Gravenhage, 1957) especially pages 17 and 56.

Most of the readers of this treatise will not be trained in linguistics. However, I do not feel that anyone interested in learning the procedures described will find the reading too difficult, even though the work was not written as a textbook. It is my intention to supplement this work by a textbook that will enable scientists and librarians to use chemical nomenclature for literature searches and for indexing without getting into the detailed understanding of organic chemical structure and theory. As a follow-up of this dissertation, work is now in progress on the completion of the lexicon of chemical morphemes. In the present work, linguistic analysis was confined primarily to acyclic chemistry while definitely establishing the feasibility of handling cyclics. To complete the linguistic analysis now requires considerable work. For example, the analyses must account for the difference in meaning of *oic acid* when it occurs with *pentanoic acid* and *benzoic acid*. This example also illustrates the futility of any syllabic approach to the study of *chemico-linguistics*.

I wish to stress that it is not necessary for the reader to wait for the appearance of the above-mentioned lexicon in order to use the algorithm (procedures) described here. This is especially true for those with training in organic chemistry, that is, have already memorized enough chemical nomenclature to carry through the simple calculations.

In closing I should like to encourage my readers to communicate with me concerning any portion of this work.

Eugene Garfield
INSTITUTE FOR SCIENTIFIC INFORMATION
Philadelphia 3, Pa.

July 17, 1961

PREFACE

This dissertation discusses, explains, and demonstrates a new algorithm for translating chemical nomenclature into molecular formulas. In order to place the study in its proper context and perspective, the historical development of nomenclature is first discussed, as well as other related aspects of the chemical information problem. The relationship of nomenclature to modern linguistic studies is then introduced. The relevance of structural linguistic procedures to the study of chemical nomenclature is shown. The methods of the linguist are illustrated by examples from chemical discourse. The algorithm is then explained, first for the human translator and then for use by a computer. Flow diagrams for the computer syntactic analysis, dictionary look-up routine, and formula calculation routine are included. The sampling procedure for testing the algorithm is explained and finally, conclusions are drawn with respect to the general validity of the method and the direction that might be taken for future research. A summary of modern chemical nomenclature practice is appended primarily for use by the reader who is not familiar with chemical nomenclature.

ABSTRACT

An algorithm for translating directly from chemical names to molecular formulas is described. The validity of the algorithm was tested both manually and by computer. Molecular formulas of several hundred randomly selected chemicals were calculated successfully, verifying the linguistic analyses and the logic of the computer program.

The algorithm for manual human translation consists of eight simple operations. The procedure enables non-chemists to compute molecular formulas quickly without drawing structural diagrams. The machine translation routine is rapid and requires a program of less than 1000 instructions. If the experimental dictionary were expanded to include low frequency morphemes, formulas for all chemical names could be handled.

The problem of chemical nomenclature is discussed in terms of the information requirements of chemists. The approach of the linguist to the problem of nomenclature is contrasted with that of the chemist. It is shown that there is only one language of chemical nomenclature though there exist many systems of nomenclature. The difficulties in syntactically analyzing *Chemical Abstracts (C. A.)* nomenclature results from *C.A.*'s ambiguous use of morphemes such as *imino*, not the use of so-called *trivial* nomenclature. The more *systematic I.U.P.A.C.* nomenclature includes idiomatic expressions but eliminates all homonymous expressions.

The structural linguist tries to describe a language compactly. While this study does not not include a complete grammatical description of chemical nomenclature, all of the basic facets of such a grammar have been studied. These linguistic studies include a morphological analysis

of the most frequently occurring segments. Approximately forty morphemes such as {*o, e, y*} and allomorphs such as *thi* and *sulf* were isolated. A list of their 200 actual co-occurrences were compiled. These studies are particularly valuable in identifying idiomatic expressions such as *diaz*, the meaning of which cannot be computed from the referential meanings of *di* and *az*. Morpheme classes are illustrated by the *bonding* morphemes (*an, en, yn, ium*, etc.) and the homologous *alkyl* morphemes *meth, eth, prop, but*, etc.

The syntactic analyses include the demonstration of transformational properties in chemical nomenclature as e.g. in *primary amines* (R–N) where *aminoRane* $\rightleftharpoons$ *Rylamine*. To complete the grammar one would have to expand the inventory of morphemes, morpheme classes, and the list of transformations. Chemical name recognition is not simply a word-for-word translation procedure. Rather the syntactic analysis required is comparable to the procedure employed by Harris, Hiz, et al (Transformations and Discourse Analysis Projects, Univ. of Pennsylvania) for normal English discourse. The structural linguistic data is supported by a summary of *I.U.P.A.C.* rules for generating chemical names.

In order to relate this study to the general problem of chemical information retrieval, the historical development of chemical nomenclature is traced from the 1892 Geneva Conference to the present. The relationship between nomenclature, notation, indexing and searching (retrieval) systems is discussed. In particular, the need for linguistic studies to solve the intellectual facet of the "retrieval" problem is discussed in contrast with the manipulative aspects which are more readily amenable to machine handling. The problem of synonymy in chemical nomenclature must be resolved if computable syntactic analyses of chemical texts are to be used for mechanized indexing. The completion of the detailed grammar of chemical nomenclature would not only permit the calculation of molecular formulas but also the generation of structural diagrams, systematic names, line notations, and other information required in machine searching systems. With suitable modifications the procedures could easily be applied to foreign nomenclature.

The field of chemico-linguistics is of interest to the organic chemist as it can improve methods for teaching nomenclature. Similarly, for the linguist chemical nomenclature is a fertile field of study. One can control the experimental conditions more easily than in normal discourse. However, conclusions can be drawn which may have more general application.

TABLE OF CONTENTS

ORGANIC CHEMICAL NOMENCLATURE – HISTORICAL BACKGROUND

INTELLECTUAL INDEXING TASKS REQUIRING STUDY

STRUCTURAL LINGUISTICS APPROACH TO CHEMICAL NOMENCLATURE

TABLE OF CONTENTS (continued)

LIST OF TABLES

ORGANIC CHEMICAL NOMENCLATURE — HISTORICAL AND BACKGROUND INFORMATION

The Contradictory Goals of Chemical Nomenclature

"It is possible in the domain of organic chemistry to give several names to the same compound. This state of affairs has on the one hand the great advantage of permitting clear expression of thought and of rendering it easier to bring out analogies in structure wherever this is useful." [*J. Am. Chem. Soc.* 55, 3905(1933)].

These remarks are quite indicative of the general state of affairs of chemical nomenclature. They are the opening sentences of the 1930 "Definitive Report of the Commission on the Reform of the Nomenclature of Organic Chemistry" (opus cited, p. 3905) and like much that is said about nomenclature, the one sentence contradicts the other.

If it is possible to name the same chemical compound in two or more different ways, does this *really* permit clear expression of thought. It depends on one's orientation. For the speaker, synonyms do indeed allow for greater freedom of expression and the ability to bring out subtleties that might otherwise be difficult to make. For the listener, such freedom of expression on the part of the speaker may result in complete loss of comprehension. To complete the round of contradictions we find in the next sentence: "But on the other hand a multiplicity of names for the same substance constitutes a serious obstacle in the preparation of indexes." (opus cited, p. 3905).

Oral Communication Versus Indexing

Thirty years ago it was not yet quite apparent to experts in chemical nomenclature that their attempts to modify prevalent nomenclature for indexing purposes actually might be making oral and written communication even *more* difficult. It is not my purpose or intention to criticize the work of these experts. The purpose of these introductory remarks is to indicate that committees on chemical nomenclature are indeed faced with the baffling dichotomy of trying to serve the purposes of oral *communication* on the one hand and the needs of *indexing* on the other. This is like trying to get people to speak the King's English in order to simplify the task of preparing dictionaries. The inability to make these two functions blend is quite obvious if one examines, briefly, the history of organic nomenclature for the past seventy-five years.

Geneva Nomenclature

Modern chemical nomenclature "officially" began in 1892 [Pictet, *Arch. sci. phy. nat.* **27**, 485-520(1892)][Tiemann, *Ber.* 26, 1595-1631(1892)] at the well known Congress of Geneva. All students of elementary organic chemistry are still taught the "Geneva" system though some teachers

may now call it the I.U.P.A.C. system. The next major revision of the Geneva system came with the 1930 Report mentioned above. Thirty-eight years later "the intent of the Geneva Congress had not been realized" i.e., Rule I enabled each chemical to be named officially so that it would "be found under only one entry in indexes and dictionaries." (opus cited, p. 3906)

I.U.P.A.C. Nomenclature

The next major report on Organic Chemical Nomenclature came almost thirty years later and is known as the 1957 Report [*J. Am. Chem. Soc.* 82,5545-84(1960)]. It is important to note that the 1957 Report contributed nothing to affect this dissertation. Most of the report is devoted to cyclic compounds. The portion of acyclic chemistry which is discussed, the hydrocarbons, does not in any way affect the linguistic aspects of my research. For that matter, as is noted below, it does not affect the basic description of organic nomenclature.

The nomenclature of so-called simple functions, i.e. substances which contain only one kind of function such as acids, alcohols, etc. are not covered in the 1957 Report. The same is true of the complex functions.

Constant Activity in Nomenclature Field

The failure of the 1957 Report to treat the entire domain of organic nomenclature does not mean that there has not been a great deal of attention devoted to chemical nomenclature during the past thirty years. On the contrary, as Austin M. Patterson noted in 1951 there were so many committees on nomenclature that it was necessary to compile a directory of them. (*Chem. Eng. News* May 28, page 2181, cited in his "Words about Words" Washington, *Amer. Chem. Soc.*, 1957). This is a collection of nomenclature columns written by Patterson for the weekly organ of the Society, *Chemical and Engineering News*.

No Basic Change

Looking at the development of organic nomenclature from the viewpoint of structural linguistics one is forced to conclude that while there are changes in the Geneva System contained in the 1930 Report, the former system is retained basically intact. Only minor details were modified.

The present situation in organic chemistry may be described by posing the following questions. If I had been ignorant of the 1930 and 1957 Reports on organic nomenclature and had compiled the list of morphemes and their corresponding syntactic rules, how accurately would this analysis describe organic nomenclature as it is used today. At least 90% of the new chemicals made each year would be recognized by a grammar based on the Geneva System. It would be an interesting study to make an *exhaustive* analysis of chemical nomenclature prior to 1892. This would determine

the basic list of morphemes available to the chemist at the Geneva conference. However, such a comparison was not germane to the particular research involved in this dissertation.

While it is true that "official" nomenclature began at the Geneva conference, examination of the 1892 Report and others (e.g. Armstrong, *Proc. Chem. Soc.* 1892,127-131) and similar examination of earlier nomenclature practice reveals that the morphology of organic chemistry not only remained essentially the same in the 1930 and 1957 Reports, (which were presumably revisions of the 1892 Geneva Report), but even the Geneva conference did not contribute any major morphological changes in organic nomenclature. The Geneva chemists simply accepted the morphological pattern already in use and codified it. In other words, a morphological analysis of organic nomenclature conducted in 1891 would have produced almost exactly the same results as an analysis conducted after the Geneva Conference in 1892.

This is not to underestimate the value of the Geneva Conference. It has served a useful function in teaching nomenclature, as there was not then available any internationally accepted system that teachers could use. However, while the teaching of organic nomenclature was not quite formalized in 1891, the terminology acquired in studying elementary organic chemistry as e.g. by using an 1890 textbook would be not significantly different than that which would be acquired in reading the same textbook in its 1920 edition in which the Geneva system is adopted.

Longest Versus Shortest Chain Structure

The Geneva Conference *did* make some significant contributions to the syntactical description of organic nomenclature, or at least to solidification of syntactical practices used by many but not as universally as was the morphology. Thus, *triethylmethane* became *3-ethylpentane*. The example of *triethylmethane* demonstrates the point well. The morphemes *tri, eth, yl, meth, an, e* were not new. Neither were the morphemes *eth, yl, pent, an, e* in *ethylpentane*. The new rules specified the selection of the latter combination of morphemes for the chemical CH_3–CH_2–$CH(CH_2$–$CH_3)$–CH_2–CH_3 by establishing the syntactical principle that the "parent" structure shall be the one which contributes the *longest* possible chain of carbon atoms. The older method of naming this chemical had an *implied* syntactic structure where one named chemicals in terms of the *shortest* chain. The same diagram can be written $(CH_3$–$CH_2)_3$–CH. There are historical reasons for this change.

Rapid Change in Syntax – Not Morphology

Early organic chemistry naturally was concerned with chemicals of simpler structure such as *methane* gas. As the knowledge of chemical structure increased, chemicals like *pentane* were easier to understand, but still the Geneva chemists could not foresee the rapid development of or-

ganic chemistry that would take place, in which it would again become necessary to modify the syntax of nomenclature but not the morphology. This would seem to be the opposite of historical development of languages where it is the morphemes which change more rapidly than syntax.

Reading Organic Chemistry as a Language

Contrary to general belief, organic chemical nomenclature is relatively simple. It is not to the credit of many teachers of organic chemistry that many students are frightened away from organic chemistry because they are confronted too early and quickly with what seem to be very complicated chemical words. Students are not taught the basic elements of organic nomenclature before they begin the formal study of the actual experimental science. This is unfortunate. One can recall that it use to be a requirement for pre-medical students to study Latin. This was really not necessary to the study of medicine. However, having removed Latin from the medical curriculum there remains a vacuum. Special preparation in the language of medicine is needed to fill this vacuum. Similarly, the special language of organic chemistry should be taught first.

Implications for Teaching Organic Chemistry

I believe there *are* implications to be drawn from this dissertation for the teaching of organic chemistry. Teaching chemistry cannot be divorced from the general problem of chemical communication. Ilowever, I cannot hope to pursue, in detail, all the derivative problems related to chemical nomenclature.

Increased Volume of Chemical Literature

As was stated in the opening paragraph, the earlier international committees on organic nomenclature tried to resolve *simultaneously* the problem of communicating and indexing chemistry. If the problem of indexing chemicals was already a problem in 1892, it is quite understandable that the emphasis on the indexing implications of nomenclature have increased. Whereas a few thousand new chemicals were prepared each year at the turn of the century, over 75,000 new chemicals were prepared by the world's chemists in 1960 alone (cf. E. Garfield, *Index Chemicus*, 1st Cumulative Index, 1961, 33.)

Notation Systems

This volume has increased the preoccupation of nomenclature experts with indexing requirements. This includes not only conventional indexing systems, but also systems which will employ machines both for listing chemicals in the conventional fashion and also for new types of machine searching The newer "nomenclature" systems, e.g. G.M. Dyson [(1947) Longmans, N.Y. 1949]and

W.J. Wiswesser (A Line Formula Chemical Notation, Crowell, N.Y., 1954) have completely discarded the semblance of English and employ completely symbolic representations. These so-called cipher or notation systems do undoubtedly simplify the problem of arraying formulas in indexes, just as notation systems simplify the problem of arraying books on a library shelf. However, just as library classification systems cannot place the book on more than one shelf at a time, using a notation system, per se, does not resolve the need to locate chemicals in more than one place in the index.

The various notation systems which have been proposed purport to avoid the pitfalls of nomenclature. None of them have been designed on the basis of a formal linguistic analysis of nomenclature. Rather, their inventors have been preoccupied with such problems as economy of notation and the ability to use the system simultaneously for the unique identification of chemical compounds as well as for generic searching. This now introduces a factor which begins to explain the background purpose of this research program.

Objectives of Linguistic Analysis

One can perform linguistic analysis with many different objectives in mind. Indeed, it is quite possible to visualize a situation in which a language might be analyzed without the linguist acquiring a speaking knowledge of that language. Similarly one can analyze nomenclature either with the idea of mastering the techniques of naming chemicals or one may be more interested in uncovering new methods of classifying chemicals. Since modern formal linguistics certainly helps one to perceive semantic as well as grammatical categories more directly than the older, more intuitive methods, (comparable to *a priori* elucidation of chemical classifications) then it is of interest to explore the possibilities of using formal structural linguistics in studying the problem of chemical information retrieval. I first discussed this possibility with Prof. Z. Harris in 1955 (E. Garfield, private communication "Structural Linguistics and Mechanical Indexing, 1955).

Information Requirements of Chemists

To completely understand the *raison d'etre* of this research, it is necessary to review some of the general information requirements of the chemist and how chemical nomenclature is related to these requirements. The organic chemist may spend years attempting to synthesize a particular chemical. In order to avoid the possibility of repeating experiments which were performed by others, he must have access to comprehensive indexes. Such indexes are typified by the *Chemical Abstracts* (C.A.) Subject and Formula Indexes (Chemical Abstracts, Columbus, Ohio).

In the C.A. indexes one can find a specific chemical by either of two methods. If one understands the C.A. system of chemical nomenclature then one can name a particular chemical in

which one is interested and look for it in the alphabetic subject index. On the other hand, if one does not have mastery of the C.A. nomenclature system one still has the option to use the Formula Index. (Incidentally, not more than a few hundred chemists in this country have a complete *mastery of the C.A. system. Three years of full-time indexing work are generally required to train* a graduate organic chemist to be an indexer for *Chemical Abstracts.)*

Formula Indexes

The Formula Index is a simple device in which each chemical is listed in alpha-numeric order according to the number of carbon and other atoms contained in it. *Ethyl alcohol (ethanol)* is listed under C_2H_6O while *acetic acid (ethanoic acid)* is $C_2H_4O_2$. By simply counting the number of carbon and other atoms in the chemical, the chemist can compute the molecular formula. With no special training he can use the formula index to find the C.A. name of the chemical in which he is interested.

I wish to make clear that these are oversimplified statements for the purpose of explanatory clarity. In actual practice one must be very cautious in calculating a molecular formula as the more complex molecules prepared today can even be difficult to depict in ideographs. This then brings up another vital question, which is, the use of structural diagrams (ideographs).

Structural Diagrams

While a chemist may frequently *not* be able to name a chemical from a structural diagram, according to the I.U.P.A.C. or C.A. systems, he *can* usually draw a diagram from a name. In order to calculate the molecular formula of a complex molecule the chemist will invariably draw its structural diagram and then proceed to add the number of carbon and other atoms. A particularly annoying aspect of working with someone else's diagram is the frequent practice of omitting some of the hydrogen atoms in the diagrams. Hydrogen atoms as such are usually of little interest to the chemist.

All existing methods of naming, indexing, coding and ciphering chemicals are based on the assumption that the chemist will first provide a structural diagram. It is important to keep this in mind when comparing methods of handling chemical information. For example, when a chemical originally reported by name is indexed by *Chemical Abstracts* the indexer will *first* draw a structural diagram. He will then proceed to rename it "systematically". More often than not, the newly assigned name will be completely incomprehensible to the chemist who first prepared the chemical. The indexer will also use the structural diagram to calculate the molecular formula which, as we have seen, is very useful to the chemist in finding a chemical in a formula index.

Molecular Formulas in Analytical Chemistry

The molecular formula also plays another important role in chemical research as it is essential in analyzing chemicals to identify them through molecular or empirical formulas. The empirical formula shows the ratio between carbon, hydrogen and other atoms. For this reason, it is generally required that the chemist report the "calculated" molecular formula of each new compound he prepares when submitting a paper to a scientific journal. It is significant that a large number of the molecular formulas reported by authors contain errors. This statement is based on my personal experience in editing the indexing of more than 100,000 new chemical compounds. Surprisingly few chemists know the "odd-even" rule which requires that the hydrogen count is an odd number if there is an odd number of other atoms present. Most of the errors are in the hydrogen count. "The calculation of correct molecular formulas requires great care and checking is justified." (E. J. Crane: "*C A Today – The Production of Chemical Abstracts*, Amer. Chem. Soc., Washington, D.C., 1958, p. 86).In this same book Dr. Crane also discusses the frequent errors found in original journal articles (opus cited p. 74).

Generic Searches

While the subject and formula indexes to *Chemical Abstracts* are designed primarily to help the chemist find a specific chemical in which he is interested, they are not especially useful when he is trying to find a chemical of related structure. Indeed, in this case the chemist may not even know the existence of a particular chemical before he begins his search. Thus he may be interested in learning whether any member of a class of chemicals has been reported in the literature as e. g. *hexanols*. Generic searching is not always practical with the conventional indexes. For this reason other methods, both manual and machine, are now extensively employed.

Chemical-Biological Coordination Center Code

The most comprehensive classification system designed for searching chemicals generically is the system of the now defunct Chemical-Biological Chemical Coordination Center of the National Research Council. This system is based primarily on the work of Prof. D. Frear of the Pennsylvania State University (CBCC Chemical Code, National Research Council, Washington, 1948.)

Modifications of CBCC Needed

The CBCC chemical code is an elaborate hierarchial system of classification based on *a priori* assumptions concerning the classes one may wish to search in large files of chemicals. While the CBCC system is quite useful, almost without exception, chemists who employ it must make modifications in particular parts of the classification schedules to differentiate more

precisely their particular chemical interests. For example, a steroid chemist would expand certain sections of the code where it is not sufficiently specific to distinguish large numbers of chemicals which might otherwise receive the same code number. This is the same problem that librarians encounter in using systems such as the Dewey Decimal System and the Library of Congress classification system.

Thus the laboratory chemist has two general requirements in searching for chemicals – the search for a *specific* chemical and the *generic* search. Turning from the chemist who is the user of indexes, what is the problem of the chemist who prepares these indexes.

The Indexer's Problem

In attempting to satisfy the information requirements of the lab chemist, the chemical indexer must deal with dozens of foreign languages in which chemical papers are written. He must also deal with the different synonym-producing-systems of naming the same chemical in each foreign language. In other words, French chemists not only have their little devices for naming chemicals, but in France, as in other countries, each chemist has certain preferences for naming chemicals in which he is a specialist.

Nomenclature Requires More Than Cooperation

The last comment may sound strange when one considers the obvious desire and willingness of chemists all over the world to cooperate in using standardized nomenclature. However, nomenclature is a problem that is far beyond the mere question of cooperation. It takes more than good intentions to resolve problems that arise from the vagaries of language. The plethora of chemical synonyms presents a formidable obstacle to the chemical indexer. If some method could be found for indexing chemical names without the many costly and enervating steps now required, a worthwhile step would have been made in documenting the literature of chemistry. This problem has great economic significance to indexer and user alike. The budget of Chemical Abstracts is over five million dollars per year.

Machine Indexing

The use of machines to perform indexing is by now no novel idea. My own investigations on the use of computers to index chemical information began in 1951 as a member of the Johns Hopkins Machine Indexing Project (cf. W.A. Himwich, H. Field, E. Garfield, J. Whittock, S.V. Larkey, Welch Medical Library Indexing Project Reports, Johns Hopkins University: Baltimore, 1951, 1953, 1955.)

Manipulative Versus Analytical Aspects of Indexing

In September of 1952, I presented an oral report on a tentative method for preparing the indexes to *Chemical Abstracts* before the American Chemical Society's Committee on C.A. Mechanization. However, most of the early work in the use of computers for scientific documentation concerned itself with the *manipulative* aspects of the problem rather than the *analytical* aspects. (cf. E. Garfield: Preparation of Printed Indexes by Machines, *Am. Documentation*, 6:68–76, 1955 and Preparation of Subject Heading Lists by Automatic Punched-Card Techniques, *J. Documentation*, 10:1–10, 1954).

In private communication to Prof. Arthur Rose, Pennsylvania State University, then chairman of the American Chemical Society Committee on C.A. Mechanization, the relationship between the problem of mechanical translation of languages and the problem of mechanical analysis of scientific literature was discussed. As the years have passed, the general awareness that the linguistic problems of indexing are far more significant than the manipulative aspects has increased. All workers in the field of information retrieval are now more conscious of the need to concentrate on problems of using computers as a substitute for the costly *intellectual* analysis required to index scientific documents by the conventional criteria as well as new criteria.

Soviet and British Nomenclature

In recent years Soviet scientists have also been devoting more attention to these problems as, for example, in the work of Tsukerman and Vladutz (cf. A.M. Tsukerman & A.P. Terentiev, Chemical Nomenclature Translation, *Proc. Intl. Conf. for Standards on a Common Language for Machine Searching and Translation*; New York, Interscience, 1961). Indeed, what it now a Soviet textbook of organic chemical nomenclature was first published in 1955 (cf. A.P. Terentiev et al, *Nomenklatura Organicheskikh Soedmionii*, Moscow, 1955) (simultaneously published in German translation as "*Vorschlage zur Nomenklature Organischer Verbindungen*", Moscow, 1955.) It is an excellent treatment of the general subject of nomenclature. There are not too many extant works to which it can be compared. Cahn's recently published work (R.S. Cahn, *An Introduction to Chemical Nomenclature*, London, 1959) is written for the lay chemist. However, as Editor of the *Journal of the Chemical Society of London*, Cahn and Cross also prepared the "*Handbook for Chemical Society Authors*", (Special Publication No. 14, London, The Chemical Society, 1960) which has many invaluable comments on I.U.P.A.C. as well as British and American nomenclature. It also gives the dates each rule was adopted.

American Nomenclature

The definitive American work on nomenclature is a publication known to most organic chemists – "*The Naming and Indexing of Chemical Compounds by Chemical Abstracts*" (Columbus,

Chemical Abstracts, 1957). The work is simply a reprint with comments of introductory remarks to the 1954 C.A. Subject Index. Neither this work nor that of Cahn can be considered to be a critique of nomenclature. That no really complete critique of chemical nomenclature is available is not surprising. This is a subject which has represented a lifetime of work for several eminent chemists among others A.M. Patterson, E.J. Crane, L.T. Cappell and the staffs of several publications in this country and abroad.

Accelerated Interest in Mechanical Analysis

The increasing availability of high-speed, high-density storage computers has now accelerated interest in the mechanical analysis of texts. It is not surprising that many individuals and teams are working simultaneously on many aspects of this problem. The possible use of computers for mechanical analysis of texts is not just an academic question involving the study of language, information theory, etc., in an academic sense, not that there can be too much research on these subjects. However, as one witnesses the growing volumes of scientific publications and the increasing difficulties of finding qualified personnel with scientific and indexing training one must be tempted to explore the full potential of the computer for every facet of indexing work. As the editor of a chemical index, I am only too well aware of the need for such assistance, even though a complete resolution of all extant problems seems now to be "futuristic". What then are the possibilities of using the computer to perform such intellectual analyses?

INTELLECTUAL INDEXING TASKS REQUIRING STUDY

Mechanical Reading Device

In the first place, one would like to have available a device for mechanically reading the words. This would avoid the costly step of manually creating a computer input in machine language. For example, one would like to index chemical papers merely by underscoring pertinent chemical names in a text. These words then would be analyzed by the computer. This was the basic premise of Frome's experiment (cf. J. Frome, U.S. Patent Office, Report No. 17, 1959).

Selective Word Recognition – Copywriter

In the work of indexing for the *Index Chemicus*, chemists must underscore pertinent chemical names and formulas. At present, there is no device available which would permit one to *selectively* "read" or "sense" printed texts, though the character recognition problem is gradually finding a solution. Large sums are now going into research on character recognition devices. However, the immediate prospect of devices which can simultaneously read the hundreds

of different typographical styles now employed is still only on the horizon. Nevertheless, a prototype "reading" unit for selectively copying words for indexing and other purposes has been invented and built by this writer and is called the COPYWRITER (cf. Fourth Annual Report, Council on Library Resources, Washington, 1961, p. 30). This machine might be modified for use in character recognition machines for selected fonts (cf. Z. S. Harris, Intl. Conf. on Scientific Information, p. 949). Since one does know the particular typographical style used by publications regularly indexed, character readers can be built to accommodate these typographical styles (cf. J. Rabinow, *Character Recognition Machines*, 1961).

Chemical Names to Structural Diagrams

Assuming now that we have obtained some form of machine input either by character recognition or by manually creating a record in machine language, what do we wish to have done with this information?

Aside from the use that is made of the structural diagram by the chemist for naming chemicals systematically, and for calculating molecular formulas, one of the primary uses of the structural diagram is for communication. The organic chemist is able to comprehend a chemical most quickly when it is presented to him in the form of a structural diagram. This type of graphic presentation is absolutely necessary because the use of systematic nomenclature is frequently either too difficult or too time consuming. While it is theoretically possible to name any chemical by the Geneva system, it must be understood that this is far from true in practice. What actually happens is that certain complex configurations are assigned either a semi-systematic or trivial name. The chemist therefore overwhelmingly prefers the use of the structural diagram. However, in order to save space journals continue to use nomenclature extensively. One would therefore like to use the computer to convert chemical names back into structural diagrams.

Drawing Diagrams by Machine

At first glance the average chemist considers computer conversion of names to diagrams an impossible task. However, this is by no means the case. It is not true either in the sense of *recognizing* and understanding the chemical name itself nor in the sense that a machine cannot "*draw*". That structural diagrams can be drawn by machine is an accomplished fact. In two separate reports Opler and Waldo have shown that structural diagrams can be drawn by a computer. (A. Opler and N. Baird, Display of Chemical Structural Formulas as Digital Computer Output, *Am. Documentation* **10**: 59–63,1958) (W. H. Waldo and M. de Backer, Printing Chemical Structures Electronically, *Proc. International Conference on Scientific Information*, National Academy of Sciences, Washington, 1959, p.711–730). In fact, the diagrams drawn by Opler's computer were so realistic, few chemists would believe that it was not a photographic projection technique until they were shown exactly

how the illusion was created on the IBM 718 output tube. This particular computer output device has a television type raster. By energizing the appropriate combination of spots, one can obtain drawings of amazing complexity. If the drawings are examined from a distance, one cannot see the spaces between the spots, thereby creating the illusion that they are line drawings. This is basically the technique used in wirephoto facsimile. One can see such patterns of dots on the front page of the daily newspaper, as it is frequently necessary to transmit photos quickly, and the size of the dots consequently must be large and more perceptible to the naked eye. If the transmission rate is slowed down, one can increase the resolution to the point where the human eye cannot easily detect the presence of the dots. There is no question that we can mechanically display and print structural diagrams by computer.

Recognizing Chemical Names by Machine

If we are capable of drawing a structural diagram by machine, then we must determine whether we can indeed find a procedure for "recognizing" a chemical name in such a way that the computer can be properly instructed to draw the correct diagram. I first began to pursue this question years ago. Could a computable procedure be found for *recognizing* chemical names and what type of analysis would be required in order to find this procedure? A further question naturally concerned the design of an experiment which could be completed in a reasonable length of time, with a reasonable chance for success.

Upon examination of the complex computer programming required to reproduce a single *known* and coded chemical on a 718 display tube, it became quite apparent that to recognize a previously unknown, uncoded chemical was not a reasonable task for one person to accomplish. Opler estimated that at least ten man-years would be required *just* to write the necessary computer programs for displaying any type of chemical diagram after suitable linguistic analyses of organic chemistry had been performed (A. Opler, Private Communication 1959). For this reason it was ascertained how much effort would be required to produce conventional line formulas as e.g.

$$C\text{-}C\text{-}N(CH_3)\text{-}C\text{-}C\text{:}C\text{:}O$$

To perform this feat, as in the case of drawing structural diagrams, this requires not only *recognition* routine, but also an extremely sophisticated *generation* routine, i.e. a procedure for generating the correct line-formula. This is further complicated by the fact that most general purpose computers do not have the typographical flexibility required for conventional line-formulas. Other methods of displaying chemicals as e.g. ciphers were also explored. A search of the literature and communication with the proponents of all well known notation systems indicated that such computer routines were not available. (G.M. Dyson and W.J. Wisswisser, Private Communication).

Calculating Molecular Formulas by Machine

Subsequently, I turned to the possibility of calculating the molecular formula. As has been stated above, the molecular formula is not only a widely used method of retrieving chemical information, it is also information that the chemist frequently needs in his laboratory work. In many situations it would not be necessary to draw a diagram if the molecular formula were available. Indeed, this is a very practical problem for every chemical publication or institution which prepares molecular formula indexes. The feasibility of preparing a program for generating molecular formulas seemed reasonable and provided a useful target for research.

While it was desirable to relate the study of finding a recognition routine to some usable output goal, the search for a recognition procedure might still have been undertaken anyhow. However, it is difficult to envision any recognition procedure which would not produce some type of usable output. Even a syntactic analysis of a sentence without regard to ultimate use does produce an output. In the case of chemical nomenclature, any output that results from a recognition routine has some value.

Having limited the scope of the output, it was then necessary to define and limit the recognition capabilities.

The Quagmire of Chemical Nomenclature

Organic chemical nomenclature is at first glance a horrible quagmire that could never be crossed by the most ambitious chemist. Naturally, the average chemist thinks first of the several million chemicals that have already been reported in the literature. There is almost an unlimited number of new chemicals that can be made. New combinations of atoms are uncovered every day. C. A. maintains a cross-reference file consisting of several hundred thousand entries. However, most people are unnecessarily discouraged by this state of affairs. It is necessary to differentiate the various facets of the problem of recognizing chemical names before one comes to the conclusion that it is a problem that is too hopeless to deal with.

There are three basic types of chemical names: (1) Trivial names, (2) Systematic names and (3) Semi-Systematic or Semi-Trivial names.

Trivial Names

The problem of handling trivial names must be dealt with in two parts: (a) names which are known prior to the computer analysis and (b) names which are entirely new. Tsukerman has properly called both types of trivial names "words-provocateurs" (opus cited, p. 4).

From the point of view of machine recognition of known trivial names there exists no problem. The storage of large dictionaries in computers is no longer a serious obstacle. With the improvement of so-called random access memory units we can expect to be able to look up items in large dictionaries quite rapidly at relatively low cost. While I would not underestimate the work involved in analyzing the thousands of trade names and other non-systematic names for chemicals, the problem of trivial names is indeed essentially trivial and of no basic interest to the linguist. This is primarily a problem of locating trade names and other synonyms by reference to standard compendia.

Legislation not a Solution

Similarly new trade names can be dealt with by non-linguistic methods. This may one day require legislative action, though it is extremely doubtful that we will see in our lifetimes the elimination of the practice of naming new chemicals biographically. You don't eliminate the use of terms like "*Richstein's Substance S*" by legislation. Rigid standards might make it very difficult for people to use such names in published journals. However, the use of trivial names or semi-trivial names is absolutely essential and *necessary* in chemistry and particularly in biochemistry. Unfortunately, the chemical structure of many chemicals is not completely known for many years. Many chemicals can only be identified by a molecular or empirical formula. The complete chemical *structure* may not be understood for many years as was the case with thousands of chemicals like *insulin, penicillin,* etc.

Systematic Names

Systematic names also fall into several categories. The word "systematic" is used very loosely to mean chemical names which are (a) named according to existing nomenclature systems or (b) named on the basis of a very prescribed list of basic terms. As the Geneva system has developed, the various commissions have tried to get chemists to rely on "systematic" nomenclature of the latter type, but this is not always easy. The I.U.P.A.C. rules as they stand today allow for so many exceptions in the selection of lexical items that it is incredible to think that all chemists will ever use it with 100% consistency. Indeed, in using CA or I.U.P.A.C. nomenclature one constantly faces the situation of having to name a chemical in a way which is completely foreign to the chemist. The rules are written primarily for the use of indexers. Consequently, the above distinction which is made by I.U.P.A.C. and by such Soviet authors as Tsukerman (opus cited) between so-called systematic and trivial names almost becomes meaningless. What is a trivial name to one chemist is a systematic name to another. If you are a steroid chemist then *androstane* is not a trivial name. It is amusing to observe that the 1957 Report (opus cited, p. 73) gives up any attempt to get chemists to name *androstane* as a derivative *cyclopentanophenanthrene*, the

more systematic description. It is equally ridiculous to call *cyclopentanophenanthrene* a systematic name when one could properly call the phenanthrene portion a derivative of *benzonaphthalene*. Once you are convinced, as I am, that the development of a truly systematic nomenclature for human communication is an impossible absurdity then distinctions between trivial versus systematic names also become absurd. If, on the other hand, one treats nomenclature linguistically chemical names can be classified as idiomatic or non-idiomatic expressions whose meanings can or cannot be computed from the meanings of the participating morphemes.

Treating Nomenclature as a Language

Most difficulties in dealing with nomenclature are due to the failure to recognize, in spite of its being a specialized jargon, that nomenclature is a sub-language of English (or whatever other language is involved). It displays many features of ordinary language. If the study of organic nomenclature is tackled as a linguistic as well as a chemical problem, then you avoid pitfalls such as the trivial-systematic dichotomy. If nomenclature is a linguistic problem then it seems reasonable to analyze the language of chemistry as you might analyze any other language. To completely describe a language is to write the grammar of that language.

Since I assumed chemical nomenclature to be a "language" with complexities or a range of complexities quite different than English or other natural languages, I was prompted to inquire how linguists might deal with such problems. I was further stimulated in this direction by the words of Bloomfield (Language, 1933) and Whorf (Language, Thought and Reality). This type of associative thought and further personal contact with linguists such as Harris inevitably focused my attention on the idea of treating organic chemical nomenclature as the structural linguist would treat a previously undescribed language.

While it was not possible for me to come to the linguistic laboratory with completely clean hands, having as a chemist acquired a general familiarity with chemical nomenclature systems, I was not uncritical of it. I have been reluctant to devote a great deal of time to the complete mastery of nomenclature because I feel that it has certain inherent limitations for communication and retrieval purposes.

In discussing organic chemical nomenclature, I have tried to indicate that as indexing problems have increased, nomenclature systems have tended to become geared more to the requirements of indexers rather than chemists or communicators. Naturally, both of these forces are constantly at work and the example I gave of the change in steroid nomenclature is one which indicates a case where the nomenclature experts had to revise systematic nomenclature to the facts as they already existed. Chemists had not followed the rules and the Commission could not overcome this fact in the in-between meetings. Between the first submission of the 1957 Report and its publication in

1960, there were over twelve thousand new steroid chemicals prepared. This is a fact from personal experience, as I examined that many steroid structures during the three years in question. In the face of such a rapid accumulation of new steroids, it is unreasonable to expect that chemists would do other than follow the principal-of-least-effort in naming chemicals. Even the layman has a good idea of what cholesterol is and it would be folly for scientific commissions to ignore the facts of natural linguistic growth. Creation of names cannot wait for the calling of annual committee meetings.

Designing Nomenclature for Machine Uses

On the other hand, if nomenclature systems can be designed both to help chemists communicate better and to index more consistently, why shouldn't nomenclature be designed so that it can be understood more easily by machines? In fact, it is not at all coincidental that elsewhere in this paper I have raised questions concerning the teaching of organic chemical nomenclature to humans. I suggest that a thorough re-examination of organic chemical nomenclature in terms of simplifying the process of analyzing chemical names by computer also would be most rewarding for teaching humans.

Certain practices are already noticeable in the naming of very complex chemical structures which appear to be accelerating this process anyhow. Chemicals are becoming so complex that chemists are finding it necessary to name them systematically but not in the I.U.P.A.C. or C A sense. This usage of existing terms does make sense to the reader and to the machine. The practice is increasing of adding substituents to the end of parent structures with intervening hyphens, without regard to the established I.U.P.A.C. rules of priority. For example, prefixes and suffixes are being used interchangeably. Most chemists could not care less whether substituents are listed in alphabetical order, by complexity, or by any other criterion. In fact, deviation from these complex ordering rules for multiple prefixes led to the formulation of a new method of filing steroids alphabetically. The system avoids absurdities which result from I.U.P.A.C.'s complex ordering rules [cf. E. Garfield, Steroid Literature Coding Project, *Chem. Literature* 12(3):6(1960)].

For example, it is the general rule in naming a chemical which has a particular function repeated to use the numerical prefix *di*. Thus one encounters *hexanediol* or more specifically *2,4-hexanediol*. If one files another chemical which is also a *hexanediol*, but which also contains an acid function as e.g. *2,4-dihydroxyhexanoic acid*, one obviously must file these two chemicals in entirely different places in an alphabetic scheme. However, the latter chemical could be called *2,4-diol-hexanoic acid* since *hydroxy* equals *ol*. Further simplification of the rules might produce *2-ol,4-ol-hexan-oic acid*. Not only is this easier to learn, it is certainly easier to analyze by machine.

Designing the Experiment

In designing the experiment and limiting its scope, I had to choose some portion of organic chemistry which was sufficiently large as to allow general conclusions to be drawn for chemical nomenclature in general. I chose acyclic chemicals as this class could be easily sub-divided if necessary. The experiment would still be reasonably complete so as to demonstrate the feasibility of tackling, by a team of linguists, chemists, and programmers, the entire domain of chemical nomenclature, especially the cyclics. The present analysis could be expanded to include and deal with more than 90% of the new compounds reported in the literature and a large percentage of the older literature by use of a relatively small number of additional morphemes such as *phen*, *benz*, *cyclo*, and other cyclic co-occurrences such as *aza*, *oxa*, etc. Thus, by a process of elimination the specific objective of my experimental program was established – to find a procedure for the machine translation of chemical names to molecular formulas.

One of the practical by-product results of this research has been to delineate a manual, algorithmic method of calculating the molecular formula of chemical names without resorting to structural diagrams. As I simulated the operations performed by the computer, based on the linguistic analysis, it became readily apparent that the procedure can be used manually. I am confident that most chemists will quickly learn and appreciate the simplicity of the method. One of the greatest values of trying to mechanize is that we are forced to look at a problem in a way that was hitherto difficult. The complete algorithm is summarized in Table V on page 30.

Another practical use of this new algorithm is found in the ability to train a non-chemist clerk to calculate a molecular formula from a chemical name.

Relationship between Nomenclature and Searching

A by-product of this study is the clearer understanding of the relationship between nomenclature and chemical searching requirements. When the computer analysis of the chemical name is completed, the parsed expression that results from the analysis could be used by the computer to perform very adequate *generic* as well as *specific* searches. If the chemists specifies the type of chemical in which he is interested in terms of morphemes instead of conventional chemical class names, generic searches become quite simple. Hence, a search for all *hexenols* becomes a search for all chemicals which contain the morpheme co-occurrence *hexen* and the morpheme *ol*. If he is interested in any six-carbon-chain-alcohol he need only specify the presence of *hex* and *ol*, where *hex* must be the carbon containing morpheme, not the multiplier morpheme as in *hexachlorooctane*.

While the computer program used in this research may be of interest to the reader (and for that reason is included here), it is only incidental to the general program of this research. The general requirements of the program, the basic approach, etc. are the pertinent factors. The specific methodology of particular computers is not of vital concern, though it is certainly an interesting exercise to work with a programmer. All of the actual Univac computer coding work was done in a relatively short time. Any large and several medium sized computers could have been used.

I personally prepared the Unityper tapes both for the input of the chemicals to be tested as well as the program. However, the actual Univac program coding was done by two University programmers, Dr. J. O'Connor and T. Angell. I wish to thank them both for this assistance. The coded Univac I program is omitted for this reason and comprises approximately 1000 code steps. However, the computer operation is described in general terms by flow diagrams in Tables VII to X.

While the study has been limited to acyclic compounds I was interested to explore just how difficult would be the transition to handling cyclic structures. A few cyclic morphemes were added to my testing procedure to simplify the selection of a random sample.

The exciting results of this side excursion over the border between the cyclic and acyclic compounds is that I have found cyclics to present no insurmountable obstacles. Certainly with sufficient, but reasonable manpower, it would be possible to resolve most of the ambiguities in the nomenclature, at least as far as calculation of molecular formulas is concerned. When we enter the realm of mechanically drawing structural diagrams then we are indeed faced with some grave problems in handling cyclics. We cannot ignore positional designations, which we can do in calculating molecular formulas. This is not because the syntactic problems of positional designations is itself difficult, which it is, but because there would appear to be no immediate solutions to the problem of resolving the use, by different chemists, of different systems of numbering well known ring systems. This would be more of a problem for older compounds published before the appearance of CA's Ring Index (Patterson, Cappell, Walker, *Ring Index*, Am. Chem. Soc., Washington, 1960).

Pattern Recognition Devices

This problem leads logically to another facet of the chemical information problem. Is it possible to find a method of "reading" structural diagrams. We have assumed all along that we would usually find our raw information in the form of printed chemical names. However, it is also true one has to deal with the printed structural diagram. Whether for the purpose of calculating a molecular formula or for naming the chemical systematically, a pattern-recognition device would be required in order to completely mechanize recognition. The National Bureau of Standards has been working on this problem using topological techniques. This is an exciting area of research, but we appear to be far from a solution to the problem.

Experiments with Cyclic Compounds

Preliminary experiments involving cyclic chemicals indicate that restricting the experiment to acyclic compounds does not affect the applicability of the procedure to cyclic structures. The greatest additional linguistic work is found, not in expanding from acyclics to cyclics, but from I.U.P.A.C. to less systematized nomenclature such as is used by *Chemical Abstracts*.

STRUCTURAL LINGUISTICS APPROACH TO CHEMICAL NOMENCLATURE

I shall outline below how a structural linguistic analysis of nomenclature differs from a non-linguistic approach. For example, the Soviet chemist Tsukerman (opus cited) uses the "syllabic" approach – a natural course for a chemist with good knowledge of nomenclature to follow. He thinks on terms of prefixes, suffixes, stems or roots, radicals, etc. On the other hand, the linguist studying nomenclature would not begin with the rule book of nomenclature, but rather with the actual discourse, the chemical names created by chemists. From the actual discourse he would discover the existing practices.

In principle, it is possible for a linguist to determine the morphemes of chemistry by interrogating an informant of that language. He can then apply the procedures of structural linguistic analysis to data obtained from the informant. The ultimate objective should be the most compact statement of the morphology.† Table I is a list of morphemes which I compiled for acyclic compounds. The word primary is used to indicate that these are the most frequently occurring – not that it is a preliminary list. In that case it would be a list of morphs.

Linguistic Forms and Their Environments

The basic approach of the structural linguist is to identify forms by examining the environments in which they occur. To obtain a description of a language one must examine a large corpus of that language. Allomorphs, morphemes, etc. are determined by a process of trial and error. Since a morpheme is a linguistic class it is essential that groups of occurrences be examined simultaneously if one is to determine that any particular sequence is or is not an occurrence of a morpheme.

†Since the phonemes of English chemical nomenclature were assumed to be the same as those used in normal discourse, it was not considered necessary to study the phonology. (There were very definite problems encountered by chemists in using Geneva nomenclature which could have been avoided if the conference had given some attention to phonetic transcription. Thus, the adoption of *yne* to differentiate acetylenes from *amines* became necessary later on. However, the phonetic identity of *ene* in *alkenes* and *ine* in *amines* is still a problem.) For the problem of translating chemical names to formulas phonology was not investigated. This does not mean that phonological studies are not germane to the problem of analyzing chemical discourse, as indeed they are. Such studies would help uncover ambiguities resulting from suprasegmental morphemes as e.g. in *dimethylphenylamine*.

In linguistics you cannot decide that a sequence is a morpheme unless you examine several utterances. Structural linguistics requires that linguistic forms be examined in *various* environments. In applying this technique to chemical nomenclature the procedure is facilitated by the existence of compendia such as Chemical Abstracts [cf. *Chemical Abstracts* 39,5867–5975(1945) for lists of frequently used radicals]. Here one finds occurrences organized by frequently occurring linguistic elements. It therefore becomes relatively simple to locate many occurrences of a particular element.

For example, in scanning a long list of chemical names you find the repetition of the segment *butyl* in names such as *butyl* chloride, *butyl*amine, di*butyl*amine, amino*butyl*decanol, *butyl*-aminohexane, etc. Preliminarily one can classify *butyl* as a morph. A *morph* is defined as a *putative* (tentative) *allomorph*. Further examination of more chemical names reveals the occurrence of *but* in *but*ane, *but*ene, *but*ynal, *but*anal, iso*but*ane, amino*but*enol, etc. In addition, one finds the occurrence of *yl* in hexy*l* chloride, hexy*l*amine, dihexy*l*amine, aminohexy*l*dodecanol, hexy*l*aminohexane, etc. On this basis the first trial, testing *butyl* to be a potential allomorph, is found to be in error. We find instead the morphs *but*, *yl*, *hex*, etc. If you ask an informant whether there is a difference in the reference meaning of *but* in each of these previous occurrences he will say there is no difference. The same will be true of *yl*. We can now proceed with further tests as to the morphemic character of *but*.

Suppose now the words *nembutal* and *nembutol* are discovered. One may call *nem* a morph. We assume that *but* in *nembutal* is a morph from the previous analyses. Then we check whether we can substitute any other morph for *nem* and we find we cannot. We also try to make a substitution for *but* in *nembutal* and we cannot. This would tend to indicate that the *but* in *nembutal* is not a morph. As additional evidence that *but* in *nembutal* is not a morph we may also ask the informant if there is a difference in the reference meaning of *but* in *nembutal* and *butane*. Should the informant not be able to express strong convictions about *but* in *nembutal* then one would rely on the formal evidence which definitely indicates that it is not the same morph as in *butane*. Thus we have dealt with the fortuitious occurrence of *but* in *nembutal*. We can now proceed with further tests as to the morphemic character of *but*.

To confirm that *but* is a morpheme we find that in most of its occurrences it can be substituted by *hex* as in *hex*ane, *hex*ene, *hex*anol, etc. In addition *but* can replace *pent* in *pentane*, *pentene*, *pentanol*, etc. We can now refer to each particular single occurrence of *but* as the morph and to the morpheme {c–c–c–c} when referring to the class of its occurrences. In this fashion we establish a preliminary list of morphemes.

Free Variation and Complementary Distribution

This list may be condensed by looking for allomorphs which occur either in free variation or in complementary distribution. In I. U. P. A. C. nomenclature there is no free variation. While

I.U.P.A.C. has eliminated free variation, it has not eliminated positional variance. We do find that *thi* and *sulf* are allomorphs of the morpheme {S}. *Thi* is in complementary distribution with *sulf*. In addition, the terminal *e* is in complementary distribution with the conjunctives *o* and *y*. These make up the morpheme {e, o, y}. *Ox* and *on* are also *allomorphs*, in complementary distribution, of the morpheme {ox, on}. *Ox* always occurs with the allomorph *o* of the preceding morpheme whereas *on* occurs with the allomorph *e*.

Co-occurrences in Systematic Organic Nomenclature

A list of co-occurrences in organic chemical nomenclature was compiled using the list of morphemes in Table I. The morphemes on this list were permuted with each other. From the total list of 1600 theoretically possible co-occurrences, 199 actual co-occurrences were determined. This was done by finding texts containing the co-occurrence or from personal knowledge of actual occurrences.

Lack of co-occurrence was further tested by using Prof. N. Rubin of the Philadelphia College of Pharmacy as an informant. We systematically went over the preliminary list of theoretical combinations. Many of the eliminations are based, not on their failure to occur in organic chemistry, but their failure to occur in acyclic compounds. Thus, combinations like *aza*, *oxa*, *thia*, *ole*, *inium*, *olium*, and *azol*, do in fact occur in chemistry, but only in cyclic structures. The classified list in Table II was compiled first. Then the alphabetic list in Table III was compiled to eliminate repetition.

TABLE I

LIST OF PRIMARY MORPHEMES FOR ACYCLIC ORGANIC CHEMISTRY

1. a	11. di	21. in	31. on**
2. acid	12. e*	22. iod	32. ox**
3. al	13. en	23. it	33. pent
4. am	14. eth	24. ium	34. sulf***
5. an	15. fluor	25. meth	35. tetr
6. at	16. hept	26. nitr	36. thi***
7. az	17. hex	27. o*	37. tri
8. brom	18. hydr	28. oct	38. y*
9. but	19. id	29. oic	39. yl
10. chlor	20. im	30. ol	40. yn

Asterisked items are allomorphs of one of the following morphemes:

* = {o, e, y} ** = {on, ox} *** = {sulf, thi}

TABLE II. CLASSIFIED LIST OF CO-OCCURRENCES

a

hepta
hexa
octa
penta
tetra

acid

acid amide
acid halide
oic acid

al

alon
anal
enal
thial
ynal

am

amat
amid
amin
amon
anam
diam
enam
sulfam
thiam
triam
ylam

an

anal
anam
ane
ano
anoic
butan
ethan
heptan
hexan
methan
octan
propan

at

ate
nitrat

at

oat
sulfat

az

azid
azin
azo
azon
azox
diaz
hydraz
nitraz

brom

bromid
bromo

but

butan
buten
butox
butyl
butyn
ylbut

chlor

chlorid
chloro

di

dial
diam
diaz
dibrom
dibut
dichlor
dien
dieth
difluor
dihept
dihex
diim
diiod
dimeth
dinitr
dioat
dioct
dioic
diol
dion
diox

di

dipent
diprop
disulf
dithi
diyl
diyn

e

ane
ate
ene
ide
ime
ine
ite
one
yne

en

buten
enal
enam
ene
eno
enoic
enol
enon
enyl
enyn
ethen
hepten
hexen
iden
octen
penten
propen
thien
trien
ylen

eth

ethan
ethen
ethox
ethyl
ethyn
yleth

fluor

fluorid
fluoro

hept

hepta
heptan
hepten
heptyl
heptyn
ylhept

hex

hexa
hexan
hexen
hexyl
hexyn
ylhex

hydr

hydrat
hydraz
hydrid
hydrox
sulfhydr

id

amid
azid
bromid
chlorid
fluorid
hydrid
ide
iden
idin
idium
ido
idox
idyn
imid
iodid
nitrid
oxid
sulfid
ylid

im

ime
imid
imin
oxim
ylim

in

amin

in

azin
ino
inyl
sulfin

iod

iodid
iodo
iodox

it

ite
nitrit
sulfit

ium

idium
onium

meth

dimeth
methan
methox
methyl
trimeth

nitr

dinitr
nitrat
nitraz
nitrid
nitrit
nitro
nitroxo
nitryl

o

ano
ato
azo
bromo
chloro
eno
fluoro
hydro
ino
iodo
ito
nitro
oat
ono

o

oxo
oyl
sulfo
thio
yno

oct

octan
octen
octyl
octyn
yloct

oic

anoic
azoic
dioic
enoic
oic acid
onoic
thioic
ynoic

ol

anol
diol
enol
ol
olic
tetrol
thiol
triol
ynol

on

amon
anon
azon
dion
enon
onium
onoic
onyl
tetron
thion
trion
ynon

ox

ethox
hydrox
idox

ox

iodox
methox
nitrox
oxid
oxim
oxo
oxy
pentox
propox
triox

pent

dipent
pentan
penten
pentox
pentyl
pentyn
tripent

sulf

disulf
sulfam
sulfhydr
sulfid
sulfin
sulfit
sulfo
sulfon

tetr

tetra
tetrol
tetron
tetrox

thi

dithi
thial
thien
thio
thioic
thiol
thion
trithi
ylthi

tri

tribut

tri

trien
trieth
trihept
trihex
trimeth
trioct
triol
trion
triox
tripent
triprop
trithi
triyn

y

oxy

yl

butyl
enyl
ethyl
methyl
nitryl
oyl
pentyl
propyl
ylam
ylbut
ylen
yleth
ylhept
ylhex
ylid
ylim
ylmeth
yloct
ylpent
ylprop
ylthi
ynyl

yn

diyn
ethyn
idyn
propyn
triyn
yne
ynol
ynon
ynyl

TABLE III. ALPHABETICAL LIST OF CO-OCCURRENCES

1. acid amide
2. acid halide
3. amat
4. amid
5. amin
6. amon
7. anal
8. anam
9. ane
10. ano
11. anoic
12. anol
13. anon
14. ate
15. ato
16. azid
17. azin
18. azo
19. azoic
20. azon
21. azox
22. bromid
23. bromo
24. butan
25. buten
26. butox
27. butyl
28. butyn
29. chlorid
30. chloro
31. dial
32. diam
33. diaz
34. dibrom
35. dibut
36. dichlor
37. dien
38. dieth
39. difluor
40. dihept
41. dihex
42. diim
43. diiod
44. dimeth
45. dinitr
46. dioct
47. dioat
48. dioic
49. diol
50. dion
51. diox
52. dipent
53. diprop
54. disulf
55. dithi
56. diyl
57. diyn
58. enal
59. enam
60. ene
61. eno
62. enoic
63. enol
64. enon
65. enyl
66. enyn
67. ethan
68. ethen
69. ethox
70. ethyl
71. ethyn
72. fluorid
73. fluoro
74. hepta
75. heptan
76. hepten
77. heptyl
78. heptyn
79. hexa
80. hexan
81. hexen
82. hexyl
83. hexyn
84. hydrat
85. hydraz
86. hydrid
87. hydro
88. hydrox
89. ide
90. iden
91. idin
92. idium
93. ido
94. idox
95. idyn
96. ime
97. imid
98. imin
99. ine
100. ino
101. inyl
102. ite
103. ito
104. iodid
105. iodo
106. iodox
107. methan
108. methox
109. methyl
110. nitrat
111. nitraz
112. nitrid
113. nitrit
114. nitro
115. nitrox
116. nitryl
117. oat
118. octa
119. octan
120. octen
121. octyl
122. octyn
123. oic acid
124. ol
125. olic
126. one
127. onium
128. ono
129. onoic
130. onyl
131. oxid
132. oxim
133. oxo
134. oxy
135. oyl
136. penta
137. pentan
138. penten
139. pentox
140. pentyl
141. pentyn
142. propan
143. propen
144. propyl
145. propyn
146. propox
147. sulfam
148. sulfat
149. sulfhydr
150. sulfid
151. sulfin
152. sulfit
153. sulfo
154. sulfon
155. tetra
156. tetrol
157. tetron
158. tetrox
159. thial
160. thiam
161. thien
162. thio
163. thioic
164. thiol
165. thion
166. tribut
167. trien
168. trieth
169. trihept
170. trihex
171. trimeth
172. trioct
173. triol
174. trion
175. triox
176. tripent
177. triprop
178. trithi
179. triyn
180. ylam
181. ylbut
182. ylen
183. yleth
184. ylhept
185. ylhex
186. ylid
187. ylim
188. ylmeth
189. yloct
190. ylpent
191. ylprop
192. ylthi
193. ynal
194. yne
195. yno
196. ynoic
197. ynol
198. ynon
199. ynyl

The Problem of Syntactic Analysis in Organic Chemical Nomenclature

> In analyzing sentences "syntactic analysis" means: a procedure for recognizing the structure of a particular sentence taken as a string of elements. To state the structure of a string is to assign its words to word classes, to divide the word class sequence into substrings and to say what combinations of substrings are admitted. (Z.S. Harris, H. Hiz et al.: Transformations and Discourse Analysis. Univ. of Penna. Computing Center Annual Report. 1960, p. 43)*

By analogy, syntactic analysis of chemical nomenclature is the procedure for recognizing the structure of a particular chemical name taken as a string of elements (morphemes).

Since chemical names are often composed of long continuous strings of morphemes uninterrupted by spaces, hyphens, or brackets, it is necessary to set up a procedure for segmenting chemical words into morphemes. In some instances the chemist does this when he uses hyphens or spaces; however, in a name like *diaminopropylaminobutylhexene* the morphemes *di, amino, prop, yl, amino, but, yl, hex, ene* must be parsed as a continuous string of alphabetic characters. It is further necessary to establish the correct bracketing relationship between adjacent morphemes as e.g. between *di* and *amino* in *diaminopropylbutylhexene* on the one hand and *bis* and *aminopropylbutyl* in *bisaminopropylbutylhexene* on the other hand. In the latter case, the morpheme *bis* has a domain of operations quite distinct from that of its allomorph di. (The reader should remember that chemical morphemes are of two kinds: those which designate calculational values as e.g. *but* = C_4 and those which designate operations performed on them such as *di* = multiply by 2.)

In a comprehensive syntactic procedure for analyzing chemical nomenclature, all bracketing will be determined algorithmically. The computer procedure described in this study does it only in part. This was done to simplify the computer programming. I.U.P.A.C. rules on the use of brackets have been interpreted to mean they are always required when there is a possibility of ambiguity. In the above mentioned case *aminopropylbutyl* would be bracketed during the preparation of the input tape. This is perfectly legitimate use of the rules and I have assumed that all means to be tested are perfectly named. In a more ambitious recognition routine we would have to include additional syntactic procedures that would identify *hexene* as the parent function.

It is significant that neither I.U.P.A.C. nor C.A. accurately prescribe the limits of *bis*. In actual practice *bis* will apply to those morphemes which can be used as substituents and the implied bracketing will end before the "parent" morpheme modified by the substituents. Thus, in the case of *bis-p-methylaminophenylhydrazone* it might refer to $=N-N-(C_6H_4-NHCH_3)_2$ or

*For a more detailed treatment see Transformations and Discourse Analysis Project No. 15. Computable Syntactic Analysis. University of Pennsylvania, Dept. of Linguistics, 1959, p. 1.

$(=N-NH-C_6H_4-NHCH_3)_2$ and parentheses become essential. At the present time there appears to be no method for resolving such ambiguity except by pre-editing as was done in this experiment. (A useful function would be served if the computer determined whether *bis* was not followed by a paren. In that event the output would indicate possible ambiguity. In this case the name would not be considered to be well-formed.)

> "The successive words of each sentence are compared with the entries in a dictionary, and each is replaced by its dictionary equivalent, i. e., the class to which it belongs (e.g. verb.) The sequence of class names which now represent the sentence is scanned for class cleavage, i.e., cases where the word may belong to two or more classes (noun and verb, for example). A program is needed to decide to which class the word belongs in its grammatical context." (Harris, Z.S., Hiz, H. et. al opus cit., p. 44)

In the case of chemical nomenclature, the problem of classification would not appear to be as complex as in normal English discourse. However, in a comprehensive syntactic analysis comparable operations would have to be performed. Otherwise we could not identify *nicotinoyl morpholine* and *pyridyl morpholinyl ketone* as synonyms. In the first case, morpholine is regarded as the parent structure. In the second case, *pyridyl morpholinyl ketone*, the ketonic function is considered the parent structure. This compound could also be regarded as a derivative of pyridine. (see p. 33)

If one seeks to recognize chemical names for the purposes of calculating from them their molecular formulas, then more elaborate forms of syntactic classifications of morphemes would not appear to be necessary. On the other hand, if the routine were designed so that one could both recognize chemical words and produce them according to I.U.P.A.C. rules, it would be very important to assign each morpheme to appropriate "syntactic categories," the sequences of which constitute well-formed chemical names. A cardinal principle of I.U.P.A.C. nomenclature is the selection of the principal functional group. A functional group is "one whose designation can be added at the end of the complete name of a compound without alteration to the name other than, sometimes, elision of terminal *e*." (R.S. Cahn, opus cited, p. 46). In this case, the choice would be quite clear. It must be named as a *ketone*, as this is the only element which is classified as a functional group.

Another important classification will be based on chain length. Hence it will be necessary to identify each member of the homologous series *meth, eth, prop, but*, etc. as such so that it will be possible to decide which of several that may appear in a name will take precedence. The principle of the *longest chain* can only be applied if one can array all members of this class which contribute chain length.

Yet another distinction is made on the basis of selecting chain lengths of greatest unsaturation. Consequently, the classification based on bonding, discussed below under *Bonding Morphemes* takes on even greater significance.

To carry the analogy further, chemical nomenclature also exhibits class cleavage i.e., cases where the morphemes may belong to two or more classes. An algorithm will therefore be required which determines for a particular grammatical context the class assignment of morphemes exhibiting class cleavage. This will be particularly true of expressions which must be classified both as regards chain length and/or functional group. Thus the common element *vinyl* (CH_2=CH–) contributes both to bonding, (unsaturation) as well as to chain length-two conflicting choices according to the circumstances.

Transformations in Organic Chemistry

The analogy between chemical names and normal sentences can be completed by showing that chemical synonyms exhibit transformational relationships similar to those exhibited by sentences. By using an appropriate notation we obtain the following transformations for the class of chemicals known as *diaryl ketones*, Ar_1–(C=O)–Ar_3, where Ar_2=Ar_1(C=O) and Ar_4=Ar_3(C=O).

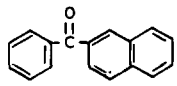
$Ar_1yl\ Ar_3\ yl\ \ ketone \rightleftarrows Ar_2oyl Ar_3ene \rightleftarrows Ar_1ylcarbonyl Ar_3ene \rightleftarrows Ar_3ylcarbonyl Ar_1ene \rightleftarrows Ar_4 Ar_1ene$

By using these transformations it is possible to generate the following list of perfectly good chemical names. Alongside each group of names is the corresponding structural diagram.

Ar_n	*A*	*B*	*C*	*D*	*E*
Ar_1 =	phen	pyridin	phen	pyridin	xyl
Ar_2 =	benz	nicotin	benz	nicotin	dimethylbenz
Ar_3 =	naphthal	morphol	morphol	naphthal	fluoren
Ar_4 =	naphthoyl	morpholenecarbonyl		naphthoyl	fluorenecarbonyl

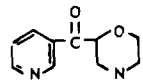
Group A

phenyl naphthyl ketone
benzoylnaphthalene
phenylcarbonylnaphthalene
naphthalylcarbonylphenene*
naphthoylphenene



Group B

pyridinyi* morpholyl* ketone
nicotinoylmorpholene*
pyridinylcarbonylmorpholene
morpholylcarbonylpyridinene*
morpholenecarbonylpyridinene

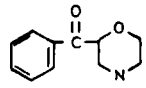


Ar_1, Ar_2, Ar_3, and Ar_4 are class designations. The synonyms for any diaryl ketone can be named by these transformation rules. One can generate well-formed names simply by specifying the values for each Ar group. This means that if one specifies the

Group C

phenyl morpholyl ketone
benzoylmorpholene

phenylcarbonylmorpholene
morpholylcarbonylphenene
morpholenecarbonylphenene



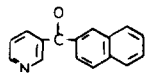
*phenene → benzene (phen → benz)
pyridinyl → pyridyl (inyl → yl)
morpholyl → morpholinyl (yl → inyl)

morpholene → morpholine (ene → ine)
pyridinene → pyridine (inene → ine)

naphthalyl → naphthyl (alyl → yl)
fluorenene → fluorene (enene → ene)

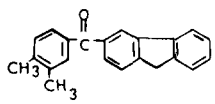
Group D

pyridinyl naphthalyl* ketone
nicotinoylnaphthalene
pyridinylcarbonylnaphthalene
naphthalylcarbonylpyridinene
naphthoylpyridinene



Group E

xylyl fluorenyl ketone
dimethylbenzoylfluorenene*
xylylcarbonylfluorenene
fluorenylcarbonylxylene
fluorenoylxylene



morpheme for Ar_1 and Ar_3 in Ar_1–(C=O)–Ar_3 a grammatically correct chemical name will be obtained by replacing Ar_1, Ar_2, etc. in the transformation equations. Prior knowledge of a correct chemical name is not required. In Table IV transformations for other chemical classes are illustrated. A thorough investigation of the transformations of chemical nomenclature would be a *sine qua non* for developing a procedure for the generation of standardized nomenclature. They are mentioned here only to complete the description of the analagous relationship that exists between syntactic analysis of normal English discourse and syntactic analysis of chemical nomenclature.

TABLE IV. TRANSFORMATIONS IN ORGANIC CHEMISTRY

Aldehydes RCH=O

R	b_n	R′	Rb_nal	formyl Rb_ne	Rb_ne carboxaldehyde
pent	an		pentanal	formyl pentane	pentane carboxaldehyde
but	en		butenal	formyl butene	butene carboxaldehyde
prop	yn		propynal	formyl propyne	propyne carboxaldehyde

Esters RCOOR

			Ryl R$b_n$oate	R$b_n$oic acid Ryl ester	Ryl Rb_ne carboxylate
eth	en	pent	ethyl pentenoate	pentenoic acid ethyl ester	ethyl pentene carboxylate
hex	an	but	hexyl butanoate	butanoic acid hexyl ester	hexyl butane carboxylate
hept	yn	prop	heptyl propynoate	propynoic acid heptyl ester	heptyl propyne carboxylate

Alcohols R–OH

		Hydroxy Rb_ne	Rb_nol
pent	en	hydroxypentene	pentenol
but	yn	hydroxybutyne	butynol

Ethers R–O–R′

			Ryl R$b_n$yl ether	
prop	yn	but	propyl butynyl ether	
hex	an	prop	hexyl propanyl ether	(propanyl = propyl)
eth	en	prop	ethyl propenyl ether	

TABLE IV. TRANSFORMATIONS IN ORGANIC CHEMISTRY (continued)

Acids RCOOH

R	b_n	R′	Rb_noic acid	Rb_n carboxylic acid
prop	en		propenoic acid	propene carboxylic acid
but	yn		butynoic acid	butyne carboxylic acid

Amines R–N

	Amino Rane	Rylamine
eth	aminoethane	ethylamine
prop	aminopropane	propylamine

The Value of Structural Linguistics for the Study of Chemical Nomenclature

The linguistic approach to the study of nomenclature provides an insight to the inconsistencies that have slowly accumulated nomenclature's natural, historical development. Linguistic analysis enables one to uncover, in advance, ambiguities that will result from the imperfect rule book of chemical nomenclature. For example, linguistic analysis indicates the occurrence of the morphemes *di*, *meth*, and *oxy* and their co-occurrence in strings such as *dimeth*, *methoxy*, and *dimethoxy*. This finding uncovers another flaw in the accepted convention of organic nomenclature and renders existing organic nomenclature far from acceptable to the machine and the human. This realization might in turn lead to a readjustment in the rules of organic nomenclature which would stipulate that all numerical prefixes be followed by parentheses. This would make the job of recognition much simpler.

It should be made clear that this study by no means purports to be an exhaustive linguistic analysis of organic nomenclature. My remarks are intended as a summary of the methods that will undoubtedly be required should a completely exhaustive study of chemical nomenclature be undertaken. In that event one would encounter many additional ambiguities in nomenclature and many new interesting morpheme classes. Expanding the scope of the linguistic analysis in this way, e.g. would bring in the cyclic chemicals which account for the majority of new chemicals prepared today. It would also introduce the complexities involved in analyzing chemical names produced not only by the I.U.P.A.C. nomenclature but also by standard British and American nomenclature. This would introduce other complexities such as variations in spelling, use of different "trivial" words, etc. (cf. T. E. R. Singer, U. S. and British Index Entries, *Searching the Chemical Literature Advances in Chemistry No. 4*, Washington: American Chemical Society, 1951.)

The Value of the Study of Chemical Nomenclature to Linguistics

In a certain sense, the domain of chemistry represents a more strictly controlled experiment for testing linguistic procedures since there are a relatively small number of parameters. It is possible, as was done in this experiment, to vary the number of parameters according to the needs of the experiment. As one gains knowledge of the language, additional morphemes and syntactical relationships can be studied so as to determine their effect on previously established knowledge. Otherwise it becomes necessary to study the language in its entirety and by the time one has even located all occurrences in the language, the natural course of human events has changed some of the relationships. This is particularly true in chemistry, where there is now a very rapid change in terminology as a result of the rapid accumulation of chemical knowledge. Certainly from the point of view of historical linguistics, one can observe changes in chemical nomenclature take place in a period of ten years that might take hundreds in normal discourse.

AN ALGORITHM FOR TRANSLATING CHEMICAL NAMES INTO MOLECULAR FORMULAS

This dissertation reports the first successful procedure for direct translation of chemical names into molecular formulas.

To test the general validity of this procedure, an experiment was designed in which certain restrictions were placed on the input and output capabilities. These restrictions were made only to facilitate experimentation with an electronic computer. As will be seen, no such restrictions are necessary when the procedure is used by human translators. Indeed, it is one of the more significant aspects of this research that it is now possible, using this procedure, to train a non-chemist to calculate, quickly and accurately, molecular formulas. This could be done by completing, for the entire domain of chemical nomenclature, the dictionary of morphemes, idioms, homonyms, etc. that has been prepared for this experiment.

The dictionary of morphemes contains, for each morpheme, the calculational value and the pertinent operations of addition and/or multiplication for that morpheme or those which precede or follow. While the experimental dictionary of morphemes is small, it is not without interest to note that these morphemes account for a large percentage of all known chemicals. The morphemes that have been eliminated are those which are ordinarily considered to be non-systematic, i.e., trivial.

The procedure was tested on a Univac 1 computer. However, any medium-sized or large computer could be similarly programmed from the general flow diagram which forms a part of this work.

TABLE V

AN ALGORITHM FOR TRANSLATING CHEMICAL NAMES TO MOLECULAR FORMULAS

SUMMARY OF OPERATIONS FOR HUMAN TRANSLATION

1. Ignore all locants (1, a, N, etc.)
2. Retain all parens.
3. Replace all morphemes by dictionary value.
4. Resolve ambiguity of any penta-octa occurrences.
5. Place + after all morphemes except multipliers.
6. If there is + at far right of parenthesized term, place it outside right paren. If there is + at far right of name, always drop it.
7. Carry out all multiplications.
8. Calculate hydrogen using hydrogen formula: $H = 2 + 2n_C + n_N - n_X - 2n_{DB}$.

Ambiguity Rules

1. You cannot have two multipliers in a row unless separated by paren.
2. If either of the next two morphemes is alkyl ending, it is not multiplier
3. If not, it is multiplier.

TABLE VI

INVENTORY OF MORPHEMES USED IN THE EXPERIMENT

			Calculation Value						
Morpheme	Meaning	Example	p	C	O	N	S	DB	I
al	O=(H)	ethan*al*	–	–	1	–	–	1	–
amide	ONH_2	methan*amide*	–	–	1	1	–	1	–
amido	$C{=}O(NH_2)$	methan*amido*propane	–	1	1	1	–	1	–
amine	NH_3	methyl*amine*	–	–	–	1	–	–	–
amino	NH_2	*amino*butanol	–	–	–	1	–	–	–
*an	–	prop*an*ol	–	–	–	–	–	–	–
*ane	–	prop*ane*	–	–	–	–	–	–	–
bis	2X	*bis*(aminopropyl) amine	2	–	–	–	–	–	–
but	C_4	*but*ane	–	4	–	–	–	–	–
di	2X	*di*aminopropane	2	–	–	–	–	–	–
*en	=	but*en*ol	–	–	–	–	–	1	–
*ene	=	*butene*	–	–	–	–	–	1	–
eth	C_2	*eth*ane	–	2	–	–	–	–	–
hept	C_7	*hept*ane	–	7	–	–	–	–	–
hepta	7X	*hepta*iodohexane	7	–	–	–	–	–	–
hex	C_6	*hex*ene	–	6	–	–	–	–	–
hexa	6X	*hexa*iodoheptane	6	–	–	–	–	–	–
hydroxy	OH	*hydroxy*ethanoic acid	–	–	1	–	–	–	–
*idene	=	butyl*iden*ehydroxyamine	–	–	–	–	–	1	–
imino	=NH	*imino*butanol	–	–	–	1	–	1	–

**bonding morpheme*

TABLE VI (cont.)

Morpheme	Meaning	Example	Calculation Value						
			p	C	O	N	S	DB	I
iodo	I–	*iodo*ethanol	–	–	–	–	–	–	1
iodoso	IO–	*iodoso*ethane	–	–	1	–	–	–	1
iodoxy	IO–O–	*iodoxy*ethane	–	–	2	–	–	–	1
meth	C_1	*meth*ane	–	1	–	–	–	–	–
nitrate	$-N=O(O_2)$	methyl*nitrate*	–	–	3	1	–	1	–
nitrile	N≡	methane*nitrile*	–	–	–	1	–	2	–
nitrilo	N≡	*nitrilo*ethanol	–	–	–	1	–	2	–
nitro	N=O(O)	*nitro*butane	–	–	2	1	–	1	–
nitroso	N=O	*nitroso*butane	–	–	1	1	–	1	–
oate	O=(O)	ethyl pentan*oate*	–	–	2	–	–	1	–
oct	C_8	*oct*ane	–	8	–	–	–	–	–
octa	8X	*octa*iodooctane	8	–	–	–	–	–	–
oic acid	O=(OH)	pentan*oic acid*	–	–	2	–	–	1	–
ol	OH	pentan*ol*	–	–	1	–	–	–	–
one	O=	pentan*one*	–	–	1	–	–	1	–
oxo	O=	*oxo*pentanoic acid	–	–	1	–	–	1	–
oxy	–O–	meth*oxy*propane	–	–	1	–	–	–	–
oyl	O=	pantan*oyl* iodide	–	–	1	–	–	1	–
pent	5	*pent*ane	–	5	–	–	–	–	–
penta	5X	*penta*chloropentane	5	–	–	–	–	–	–
peroxide	–O–O	ethylmethyl *peroxide*	–	–	2	–	–	–	–
prop	C_3	*prop*yne	–	3	–	–	–	–	–
sulfate	$-O-SO_2-O$	methyl *sulfate*	–	–	4	–	1	–	–
sulfino	HSO_2-	*sulfino*propanoic acid	–	–	2	–	1	–	–
sulfinyl	–SO–	ethyl*sulfinyl*propane	–	–	1	–	1	–	–
sulfo	HSO_3	*sulfo*propanoic acid	–	–	3	–	1	–	–
sulfonyl	$-SO_2-$	methyl*sulfonyl*butane	–	–	2	–	1	–	–
tetra	4X	*tetra*iodobutane	4	–	–	–	–	–	–
tetrakis	4X	*tetrakis*(ethylamino)	4	–	–	–	–	–	–
thial	S=(H)	ethane*thial*	–	–	–	–	1	1	–
thio	–S–	methyl*thio*ethane	–	–	–	–	1	–	–
thiol	–SH	ethane*thiol*	–	–	–	–	1	–	–
thione	S=	propane*thione*	–	–	–	–	1	1	–
tri	3X	*tri*iodopropane	3	–	–	–	–	–	–
tris	3X	*tris*(aminopropyl)amine	3	–	–	–	–	–	–
*yl	–	butyl*amine*	–	–	–	–	–	–	–
*ylene	––	ethyl*ene*diamine	–	–	–	–	–	–	–
yn	≡	butyn*al*	–	–	–	–	–	2	–
yne	≡	buty*ne*	–	–	–	–	–	2	–

**bonding morpheme*

Generalized Expression for the Molecular Formula

The result of my investigating the requirements for such an algorithm is the following simple generalized expression for a molecular formula in terms of morphemic analysis of its chemical name,

$$(1) \qquad m.f. = \sum_{p=1}^{j} p_j M_{i_n} + H$$

where P_j is the number of occurrences of morpheme M_{i_n}, i is the element (e.g. carbon, oxygen, nitrogen, etc.) and n is the number of occurrences of i in M. For chemicals which contain only carbon and hydrogen (Hydrocarbons) this expression becomes

$$(2) \qquad \sum_{p=1}^{J} p_j M_{c_n} + H$$

For chemicals containing the elements carbon, oxygen, nitrogen sulfur, and halogen the expression can be expanded as follows:

$$(3) \qquad m.f. = \sum p_j M_{c_n} + \sum p_j M_{O_n} + \sum p_j M_{N_n} + \sum p_j M_{S_n} + \sum p_j M_{X_n} + H$$

This expression covers all chemicals tested in this experiment.

Each of the terms in this latter expression can be expanded, as in the case of morphemes relating to carbon as follows:

$$(4) \qquad \sum p_j M_{C_n} = p_1 M_{C_1} + p_2 M_{C_2} + p_3 M_{C_3} + \cdots p_j M_{C_\infty}$$

where M_{C_1} is the morpheme *meth*, M_{C_2} is the morpheme *eth* and all the other terms are the members of the homologous series C_1, C_2, C_3, ... C_∞. Each of the other terms in equation (4) is the summation of all morphemes which contribute to the value of that particular atomic element.

The value for hydrogen is found from the following expression

$$(5) \qquad H = 2\left[\sum M_C - \sum M_{DB} + 1\right] + \sum M_N - \sum M_X$$

M_{DB} is the special class of morphemes which contribute double bonds, and cyclics as e.g. *an*, *en*, *yn*, and *cyclo*.

Soffer's Equation for Molecular Formula

This expression is derived in part from Soffer's generalized expression for the molecular formula in terms of cyclic elements of structure. (M.D. Soffer, *Science*, 127:880,1958).

(6) $$p = 1 + 1/2(2n_C + n_N - n_{H,X})$$

However, Soffer's equation does not take into account such elements as oxygen and sulfur, nor does it provide for chemicals such as quaternary ammonium compounds in a direct fashion. All such compounds are covered by the generalized expression pM_1. The case of quaternary compounds is particularly interesting, as its main morpheme constituent *ium* is classified by its DB value together with *en* and *yn*. *All of these* morphemes are 'bonding' morphemes. This is reasonable as in quaternary ammonium compounds nitrogen is in a pentavalent state and thereby contributes the equivalent of a double bond to trivalent nitrogen. For this reason, its DB value is minus one (-1).

Only One Language of Chemical Nomenclature

Aside from the utility of the algorithm for calculating molecular formulas, it is important to note that there really exists only one language of organic chemistry. It is a sub-language of English, but in spite of all the different "systems" available for naming chemicals, resulting in many synonyms for the same specific chemical, all of these systems draw on the same basic dictionary of morphemes. Two chemists may name the same chemical differently, but they will also be able to reconstruct the structural diagram of the chemical, and from it the molecular formula, with little or no difficulty. Upon cursory examination the chemical *2-(nicotinoyl)morpholine* might not appear to be the same as *3-pyridyl 2-morpholinyl ketone*, but drawing the structure of each, and calculating the formula would show that they are synonyms. Since there is in fact only one language involved, not several, the algorithm works regardless of the system used. It works equally well for Chemical Abstracts nomenclature as for I.U.P.A.C. nomenclature.

To illustrate the use of the algorithm a series of examples of increasing complexity are discussed. The first will illustrate the dictionary look-up routine, the second and third the use of multipliers and parenthesized expressions, the fourth a chemical requiring the use of an ambiguity-resolving routine. It is particularly interesting to observe that much of the complexity of computer programs for this type of analysis is due to the intricate steps required by the machine to recognize and deal with ambiguity. The human translator combines the ambiguity-resolving routine with the dictionary look-up routine quite easily.

First Example

As a first example consider the simple chemical name *methylaminoethane* in which there are no parenthesized terms, no positional designations (locants) or multiplier morphemes (coefficients).

Methylaminoethane is analyzed morphemically by the human translator as follows — *meth, yl, amin, o, eth, an, e.* Each morpheme is assigned the following meaning by reference to the dictionary. Since these are the most frequently occurring morphemes in the language they are memorized in the first few minutes.

meth = C
yl = +
amin = N
o = +
eth = 2C
e = +

By the process of simple addition one obtains the partially complete molecular formula as 3C + N. When written in the conventional chemical subscript notation this becomes C_3N. It now remains to calculate the hydrogen.

H = 2 + 2(3) + 1 −O −2(O) = 9 The complete formula is C_3H_9N

Second Example

As a second example let us consider the chemical

(3-(diethylamino)propyl)ethyl-3-amino-1,4-butanedioic acid

By a similar morphemic analysis this becomes

(O−[2(2C) + N] + 3C) + 2C + O + N + O + 4C + O + 2(2ϕ + DB) ϕ = oxygen

(7C + N) + 6C + N + 4ϕ + 2DB = 13C + 2N + 2DB = $C_{13}N_2O_2$ + 2DB

and where H = 2 + 2(13) + 2 − O −2(2) = 26 Final m.f. = $C_{13}H_{26}N_2O_4$

Third Example

As a third example consider *bis(bis[diethylamino]propylamino)butane.*

2[2(2[2C] + N) + 3C + N] + 4C + O

2[2(4C + N) + 3C + N] + 4C

2(8C + 2N + 3C + N) + 4C

16C + 4N + 6C + 2N + 4C = 26C + 6N = $C_{26}N_6$

H = 2 + 2(26) + 6 −O − O = 60 and the m.f. = $C_{26}H_{60}N_6$

Fourth Example

Finally, consider the example of *hexanitrohexatriene.*

6(N + 2ϕ + DB) + 6C + 3 DB

6N + 12ϕ + 6DB + 6C + 3DB = 6C + 6N + 12ϕ + 9DB = $C_6N_6O_{12}$ + 9DB

H = 2 + 2(6) + 6 − O − 2(9) = 2 and m.f. = $C_6H_2N_6O_{12}$

In this particular case the morphemic analysis is not as straightforward since there are several potentially ambiguous morpheme combinations.

Ambiguity and Principal of the Longest Match

The algorithm must account for the fact that the *hexan* in *hexanitro* is not the same as the *hexan* in a compound such as *nitrohexane* or for that matter the *hexane* buried in *hexatriene*. In the latter case the *hexa* in *hexatriene* is not the multiplier found in *hexanitro*. These ambiguities are resolved by a simple ambiguity-resolving sub-routine for the morphemes like *hex* (called *pent-oct* group in experiment). This consists of testing either one and/or two of the morphemes to the right of the ambigous *pent-oct* morpheme as to whether it is an alkyl ending (as e.g. *an, en),* a multiplier-morpheme (as e.g. *tri*) or a morpheme such as *nitro*. In order to understand how the computer procedure differentiates the *hexan* in *hexanitro*, it is necessary to explain the principal of the longest match which is used in the entire recognition procedure for assigning dictionary values to the morphemes. Since the human translator *learns*, he has no difficulty in making the differentiation.

In the experiment, it was found that the longest morpheme in the dictionary was eight letters long. For this reason, matching consists of examining the last eight letters of a chemical name first. In an expanded coverage of chemical nomenclature more letters would be matched as e. g., a morpheme such as *hentriacont*, meaning a thirty-one carbon chain. Consequently, in the example above, *hexanitrohexatriene*, the characters *xatriene* would be examined first. Since no match would be found for this combination of letters, the test would be continued with *atriene*, which again would find no match. There would be no match until *ene* was reached, at which point the last three letters of the name would be stripped and the procedure would continue with *ohexatri*. By a similar procedure, a match would be found for *tri*. Then we would match against *itrohexa* and we would find a match for *hexa*. (To simplify the procedure both *hex* and *hexa* are stored in the dictionary.) Simultaneously the *pent-oct* ambiguity-resolving routine would be called for, as each morpheme is always checked for membership in this list. The correct value of *hexa* in *itrohexa* having been determined, we would then move on to *exantro*, where we would encounter a match for *nitro*, leaving as the final residue, *hexa* which, of course, would go through the same ambiguity-resolving routine as the previous occurrence of this morpheme.

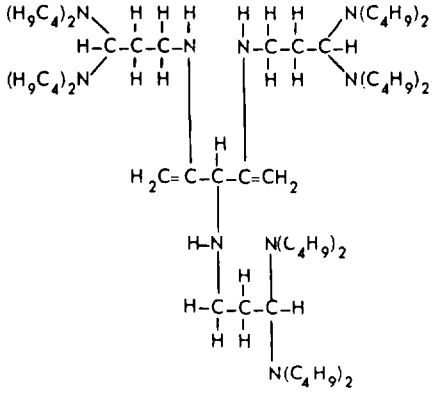
For the human translator, this procedure is by no means as complex, as one can readily perceive that *hexa* is followed by the very common morpheme *nitro* and subsequently by *tri*.

While the reader can apply the algorithm with no difficulty without a computer, the computer program may not be self-evident without reference to a specific example. For this reason, another example has been chosen which will test all of the steps in the program, including the general recognition program, dictionary look-up routine, *pent-oct* ambiguity- resolving routine, and formula

calculation routine. In order to test all boxes in the calculation routine it is necessary to select a chemical with several parenthesized expressions, i.e. nested parentheses.

Fifth Example -- Human Procedure

Consider the chemical *2,3,4-tris[3-bis(dibutylamino)propylamino]pentadiene-1,4*
Off computer the algorithm for this compound results simply in $3[2(2\ C_4 + N) + C_3 + N] + C_5 + 2$ DB. Carrying out the simple multiplications and additions gives a partial molecular formula of $C_{62}N_9 + DB_2$ and $H = 2 + 2(62) + 9 - 2(2) = 131$. m.f. = $C_{62}H_{131}N_9$. The structural diagram of this chemical is also shown to indicate how time-consuming it can be to go through the procedure of drawing such a diagram in order to calculate the molecular formula.



Fifth Example -- Computer Procedure

The computer procedure for analyzing the same compound is given below. Parenthetical remarks are made to help explain some of the details which would apply to all chemicals. The entire chemical name is punched on an IBM card or typed directly on a Unityper typewriter. The tape or card is then read into the main computer and immediately placed in a working storage unit. Working from right to left each character in the name is brought into the computer register one at a time and processed one at a time. The character in process at any instant is referred to as the *current character.*

Ignorability not Obvious Discovery

The first part of filtering each character consists of the test for 'ignorability', i.e. is it a character which cannot enter any look-up or other operations that will contribute to the molecular

formula. It is worth noting that ignorability of positional terms, i.e. locants in chemical names was no obvious discovery and had to be carefully checked for validity.

Current Character Processing

Since the first current character in processing our *pentadiene* example is an *e*, it is not ignorable. It will therefore not be possible to discuss how ignorable characters are handled until later in this example. Since the *e* is not ignorable it is then tested for being a paren and since it is not it is placed in a special storage unit called alpha storage. Immediately we ask whether there are eight characters in the alpha storage; since there aren't, we then test whether we have a sentinel character which signifies the end-of-name. In this experiment, the ampersand symbol was used for this sentinel.

Since we have not reached the end-of-name, the next character is taken out of working storage and processed in exactly the same way. This will continue, in this case, until we do have eight characters in the alpha storage *(ntadiene)*. At this point, we will process the alpha storage, initiating the dictionary match or look-up routine.

Dictionary Match Routine

The dictionary match routine will compare the contents of the alpha storage with the dictionary and will find a match for *ene*. Since this morpheme is not on the pent-oct list the morpheme *ene* will be placed in a special calculation and morpheme storage area along with its appropriate meaning. In this case it will be DB_1. The alpha storage will now be asked whether it is empty. Since it is not, all of the characters in alpha storage that remain will be shifted to the far right leaving *ntadi*. A match will be found for the morpheme *di* and it, too, will then be stored in calculation area. Numerical *multipliers* have a special code digit which is used during the formula calculation routine to differentiate them from *adders*.

Fully Processing Alpha Storage

The alpha storage is now shifted again. This time, when a match is sought for *nta*, there will be no such morpheme. Therefore, current character processing will continue until the first right paren is encountered. This paren will then cause the computer to check if alpha store is empty. Since it is not, the paren will be placed in a paren storage and the contents of the alpha storage will be *fully* processed which means that whatever characters remain in alpha storage must be one or more complete morphemes. In this case *penta* remains in alpha storage and it will go through the dictionary match routine. Since it is on the pent-oct list, it will also go through the pent-oct ambiguity-resolving routine.

Pent-Oct Ambiguity-Resolving Routine

Since the morpheme preceding *penta* is not an alkyl ending, the procedure then determines whether it is a numerical prefix. Since *di* is a numerical prefix, it is determined whether the next morpheme is an alkyl ending. Since *ene* is such an ending, *penta* will be stored in calculation area as would *pentane*, i.e. as a C_5 rather than as a multiplier. The ambiguity has been resolved. Current character processing is now resumed.

The eight characters *pylamino* will go into alpha storage and *amino* will be matched and placed in calculation area. Processing will continue and *yl* will also be matched. Processing will continue until the next right paren is encountered, at which point *prop* will be found in alpha storage, fully processed, and the paren will also be stored in the calculation area as a full word, since the alpha store will have been found to be empty. This was also done with the previous right paren when *penta* was processed. The procedure will continue similarly with *dibutylamino*, until the next paren (a left paren) is encountered. *Bis* will then be processed as a morpheme, the hyphen will be ignorable, as will the 3 and the second left paren will be encountered and placed in the calculation area. *Tris* will then be processed and the remaining characters ignored. When the end-of-name character is encountered, the formula calculation routine will be initiated. Determining whether a character is ignorable is done by a dictionary sub-routine, in which the computer compares each current character with a complete list of ignorable characters consisting of the integers 1 to 8, hyphen, comma, prime, and colon. The presence of an ignorable character will always indicate the beginning or the ending of a portion of the name which can be processed independently of the other portions.

Computer Calculation Routine

The calculation storage area of the computer now contains the following sixteen calculation words. Each morpheme is followed by its appropriate additive or multiplicative value. Note that parens also stored as separate calculation words.

Word	*Value*	*Word*	*Value*	*Word*	*Value*	*Word*	*Value*
1. tris	3(9)	5. di	2(9)	9.)	---	13.)	---
2. (	---	6. but	C_4	10. prop	C_3	14. penta	C_5
3. bis	2(9)	7. yl	---	11. yl	---	15. di	2(9)
4. (	---	8. amino	N	12. amino	N	16. ene	DB

The first portion of the calculation routine disposes of parentheses and multiplying operations. The first word *tris* is a multiplier, so it is then determined whether the next word is a left paren, which it is. The computer now starts counting left and right parens. We again ask if the next word is a paren. Since it is not, but it is a multiplier, *bis*, multiplication is not yet carried out.

Since the next word is a left paren, the count of left parens will increase to two. However, since the registers for left and right parens are not yet equal, the next word is examined. Since *di* is not a paren, but is a multiplier, it, too, will be ignored. The next word is *but*. Since it is not a numerical prefix, it will be multiplied by the multiplier *tris*. The same will occur for *yl*(7), *amino*(8), *prop*(10), *yl*(11), and *amino*(12) as they are all contained within the parens covered by *tris*(1).

When the right paren following the last *amino*(12) is encountered, the left and right paren registers will be equal. This will signal computer to return to the word immediately following the first left paren – *bis*(3). A similar process will now be followed which will result in multiplying *but*(6), *yl*(7) and *amino*(8) by two. When the paren following the first *amino*(8) is encountered, the computer will be referred back to the first *di*(5). Since it is a multiplier, is not followed by a paren, *but*(6) will be multiplied by two.

Before proceeding, the computer checks whether the last word in calculation area has been reached. Since it has not, *yl*(7) will be processed and ignored as will *amino*(8), right paren(9), *prop*(10), *yl*(11), *amino*(12), right paren (13), and *penta*(14) which had been found, during the ambiguity-resolving routine, to be C_5.

Since *di*(15) is a multiplier the morpheme *ene*(16) is multiplied by two. Since it is the last calculation word, the paren and multiplication operations are completed. All parens and multiplier calculation words are now replaced by zeros. The computer then adds the contents of these registers which now looks as follows:

	Word	*Value*		*Word*	*Value*		*Word*	*Value*		*Word*	*Value*
1.	tris	000	5.	di	000	9.	)	000	13.	)	000
2.	(	000	6.	but	$C_4x3x2x2=C_{48}$	10.	prop	$C_3x3=C_9$	14.	penta	C_5
3.	bis	000	7.	yl	000	11.	yl	000	15.	di	000
4.	(	000	8.	amino	$Nx3x2=N_6$	12.	amino	$Nx3=N_3$	16.	ene	$DBx2=DB_2$

The totals are taken and give a partial molecular formula of $C_{62}N_9DB_2$. The hydrogen calculation is performed using the equation $2 + 2n_C + n_N - n_X - 2n_{DB}$. In this case it is 131 giving a final formula of $C_{62}H_{131}N_9$.

The computer will now test for experimental purposes whether the calculated formula agrees with the formula calculated manually and stored with the original data.

Hydrogen Calculation

The calculation of hydrogen is by no means a simple straightforward or obvious task. There are two ways of solving the problem. There is the method described in this dissertation which

derives form Soffer's formula and there is the standard procedure used by chemists. To give the reader an idea of the difficulties of using the conventional method, he is referred to the complex chemical diagram shown on page 36, where the fifth example is discussed. It is obvious that the brute force method of counting 131 hydrogen atoms is likely to generate errors. To duplicate the brute force method of calculating hydrogen by an algorithm is not only difficult but also uneconomic in terms of computer time.

The assignment of computational values (semantic mapping) to a relatively small list of morphemes which also accounts for hydrogen, would at first glance, appear to be a rather trivial task. However, here one must depart from morphology and take into consideration the rules of chemical bond formation. For example, the term *methyl* consists of two morphemes *methyl* and *yl*. This is one of the most commonly occurring terms in organic chemistry and has a calculational value of CH_3. It is invariably CH_3. On the other hand, *propane* is $CH_3CH_2CH_3$. However, *methylpropane* is not merely the summation of the values for methyl and propane. In adding the *methyl* group, one must replace one of the hydrogen atoms on the *propane* nucleus giving a structure $CH_3CH(CH_3)CH_3$ more commonly called isobutane. If in compiling a dictionary of morphemes, we assign the values usually associated with the morpheme, then we must incorporate very sophisticated rules based on a knowledge of chemical formation. The problem increases in complexity when dealing with names containing morphemes such as *oate*, where a chemical reaction is implied as between an acid and an alcohol to form an ester. For example, the simple chemical *ethyl ethanoate (ethyl acetate)* is not the addition of $C_2H_5+C_2H_6+O_2$. The formula for this chemical is $C_4H_8O_2$ since an ester is formed from the combination of an alcohol and an acid with the elimination of a molecule of water.

The linguist is prompted to ask whether one has the right to include hydrogen value in the semantic mapping of morphemes such as *meth*. The morpheme *meth* will always contribute one carbon atom to the molecular formula, but it does not always contribute three atoms of hydrogen. It is not at all obvious, even to the chemist, how one resolves the problem of hydrogen calculation. It is well known that the number of hydrogen atoms in a saturated hydrocarbon is derived from the relation $2N_C + 2$ where N_C is the number of carbon atoms. However, the average chemist has no systematic method of quickly solving for hydrogen.

Soffer (opus cited) provides a more sophisticated statement of the relationship between the number of cyclic configurations in a chemical and its molecular formula. I had previously used Soffer's formula in checking the accuracy of several thousand formulas. However, it did not occur to me immediately that it could be modified and used as a means for obtaining the hydrogen value directly. It was observed that each of the terms in Soffer's equation could be replaced by a term representing a morpheme, i.e. a group of allomorphs, particularly the "bonding" morphemes contributing to 'cyclic' configuration. Then it was possible to simplify the syntactic rules for each morpheme.

The value of this approach is more apparent if one considers an example in which hydrogen is determined by the previous method of first identifying the 'parent' structure in a chemical name. The parent morpheme is frequently an alkane ending such as *ane*. The chemical *4-hydroxy-3-heptanone* is derived from *heptane*. You calculate its molecular formula by starting with C_7H_{16}, the molecular formula of heptane. The morpheme *one* adds an oxygen atom and subtracts two hydrogen atoms.

For *hydroxy* you add another atom of oxygen. *Hydroxy* contains one additional hydrogen atom, but this is balanced by the loss of one H atom in adding the *hydroxyl* substituent. This procedure works quite well for chemicals with straightforward substitution of one functional group for hydrogen. However, it breaks down in more complex cases. By confining one's dictionary to morphemes in which hydrogen is excluded and calculated after all other calculations are performed, a more straightforward procedure is possible.

Thus, the assignment of 'meaning' is conditioned by the syntactic methods that are employed for analyzing the chemical name and for generating the correct molecular formula. However, once the new approach is chosen, one must analyze each morpheme a little more closely. It is not sufficient to know that nitro is NO_2. It is necessary to learn that it is one nitrogen atom attached to two oxygen atoms, in which, one of the attachments is by a double bond. The presence of this double bond affects the total hydrogen content of the molecule. It therefore must be recorded in the dictionary along with the remaining semantic information.

Having recorded the semantic value of each morpheme, it is further necessary to provide rules for distinguishing between the homonyms which occur in systematic nomenclature. Thus, there is a class of numerical prefixes which unfortunately are ambiguous with morphemes for alkanes. For example, *pent* may be additive, as in a chain of five carbon atoms, such as pentatriene or it may be a multiplier as in pentachlorohexane. This situation is not unlike the problem of syntactic analysis of English text, in which, one finds two words in a sentence which are part of the same verb, but are separated by an intervening word, e.g. a split infinitive.

TABLE VII. GENERAL PROGRAM FOR CHEMICAL NAME RECOGNITION

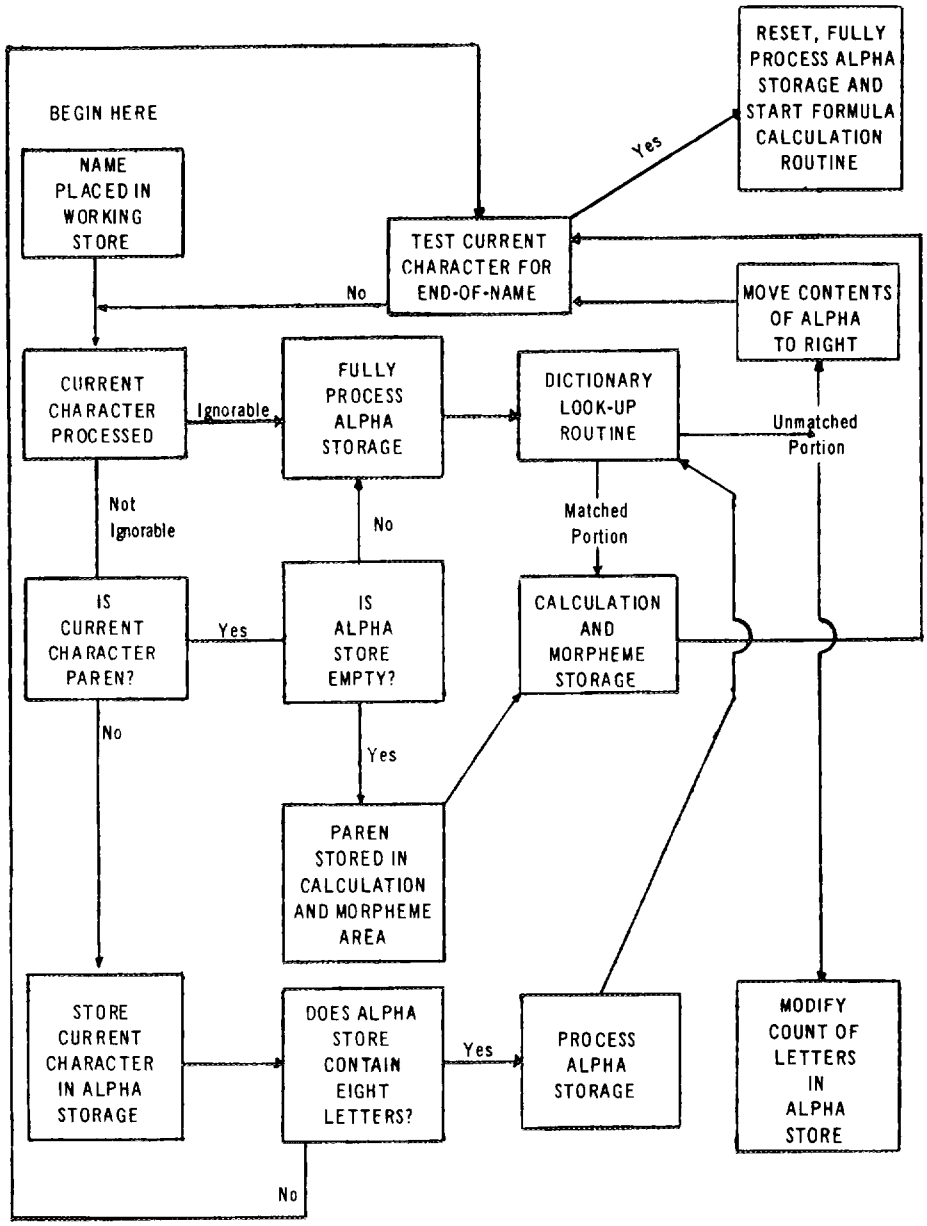


TABLE VII

CHEMICAL NOMENCLATURE ANALYSIS

COMPUTER CALCULATION OF MOLECULAR FORMULAS

GENERAL PROGRAM DESCRIPTION

1. Chemical name is typed on Unityper. Only chemical names of sixty characters or less are allowed, to simplify programming. Sixty characters are stored in five Univac words.

2. Chemical name is placed in working storage. (Left-to-right in the name is equivalent to top-to-bottom in storage.)

3. Processing of name starts with bottom character in working storage, i.e. character on the far right of chemical name.

4. Determine whether the current character is ignorable, i.e., a dash (hyphen), number, prime, comma, or delta (space).

5. If it is, then ignore it and fully process* contents of alpha storage.

6. If it is not ignorable character, determine if current character is paren.

7. If current character is a paren, store it in calculation area of storage, unless alpha storage already contains something, in which case, store the paren in paren storage and fully process* alpha storage.

8. If it is not a paren and also not ignorable, then store it in alpha storage. Continue processing until eight characters are stored in alpha storage. This is determined by counting characters as they go into alpha storage.

9. Find a "match" for the contents of the alpha storage, i.e. from the morpheme dictionary, look up value of morpheme in alpha storage. This might be the entire eight letters or just two letters, but no less than two letters, otherwise there is error signal.

10. When the match is found, enter the calculation value of the morpheme in the next available storage location of the calculation storage and the morpheme itself in the morpheme area.

11. Move any remaining unmatched portion to the far right in alpha storage. At the same time this will change the count of the number of letters in alpha storage.

*Fully process alpha storage means that whatever alphabetic characters are in alpha storage will be examined so as to identify the morpheme(s) involved. After finding a match for the right end of alpha storage the remainder will be shifted and similarly processed. However, "fully" process cannot be used if alpha storage processing was started as a result of 8 count.

2. Keep on examining more characters in name until there are again eight characters in alpha storage.

3. Continue the process until all characters have been placed in storage. When end-of-name signal (&) is encountered, computer will know that processing of all characters has been completed.

TABLE VIII. CHEMICAL NOMENCLATURE ANALYSIS

DICTIONARY LOOK-UP ROUTINE

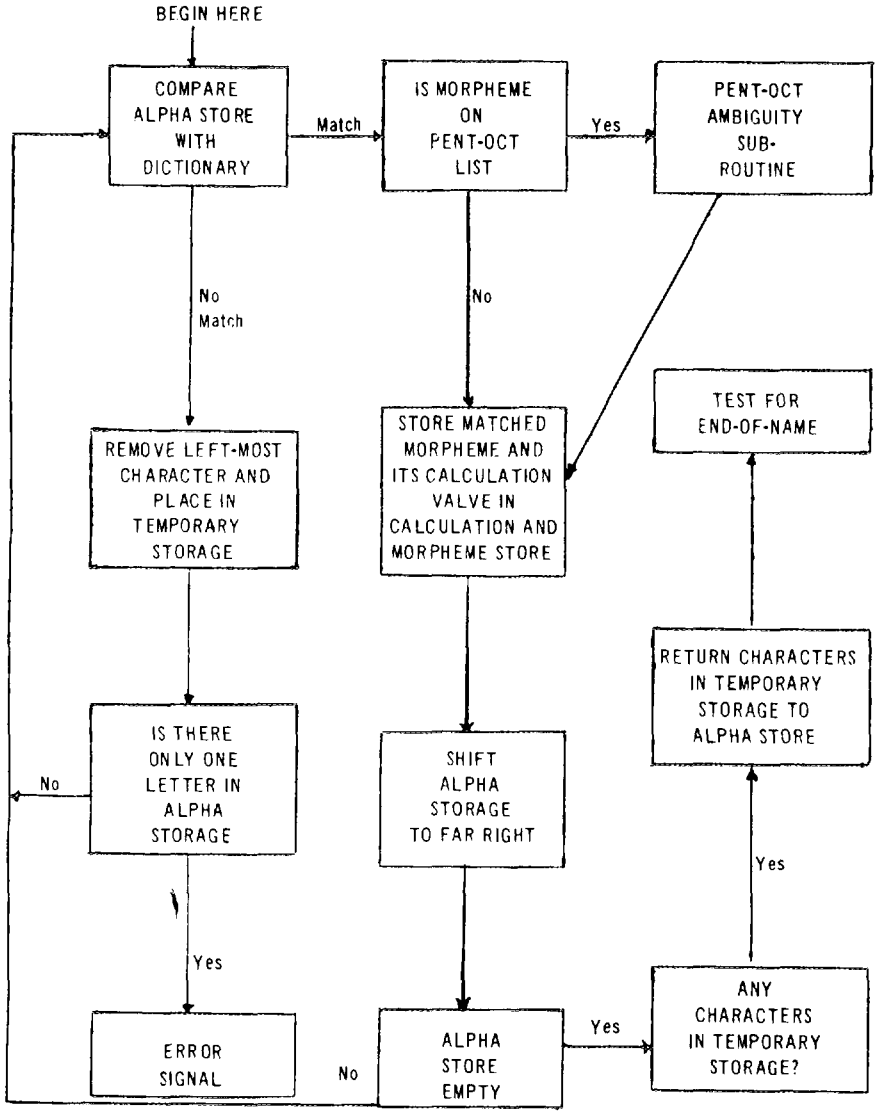


TABLE VIII

CHEMICAL NOMENCLATURE ANALYSIS

DICTIONARY LOOK-UP ROUTINE

1. The longest morpheme match is looked for first. The characters in alpha storage are compared to all morphemes in dictionary.

2. If no match is found, left-most character is dropped and matching process begins again. In this way *thial* is matched before al.

3. Before matched morpheme is stored in calculation area, it is checked for being in pent-oct group of homonyms.

4. If the morpheme is found to be in pent-oct group, then a special ambiguity-resolving routine is initiated.

5. If morpheme is not pent-oct, it is placed in calculation and morpheme storage.

6. If alpha store is not empty, it is shifted to far right and process begins over.

TABLE IX. PENT-OCT AMBIGUITY RESOLVING ROUTINE

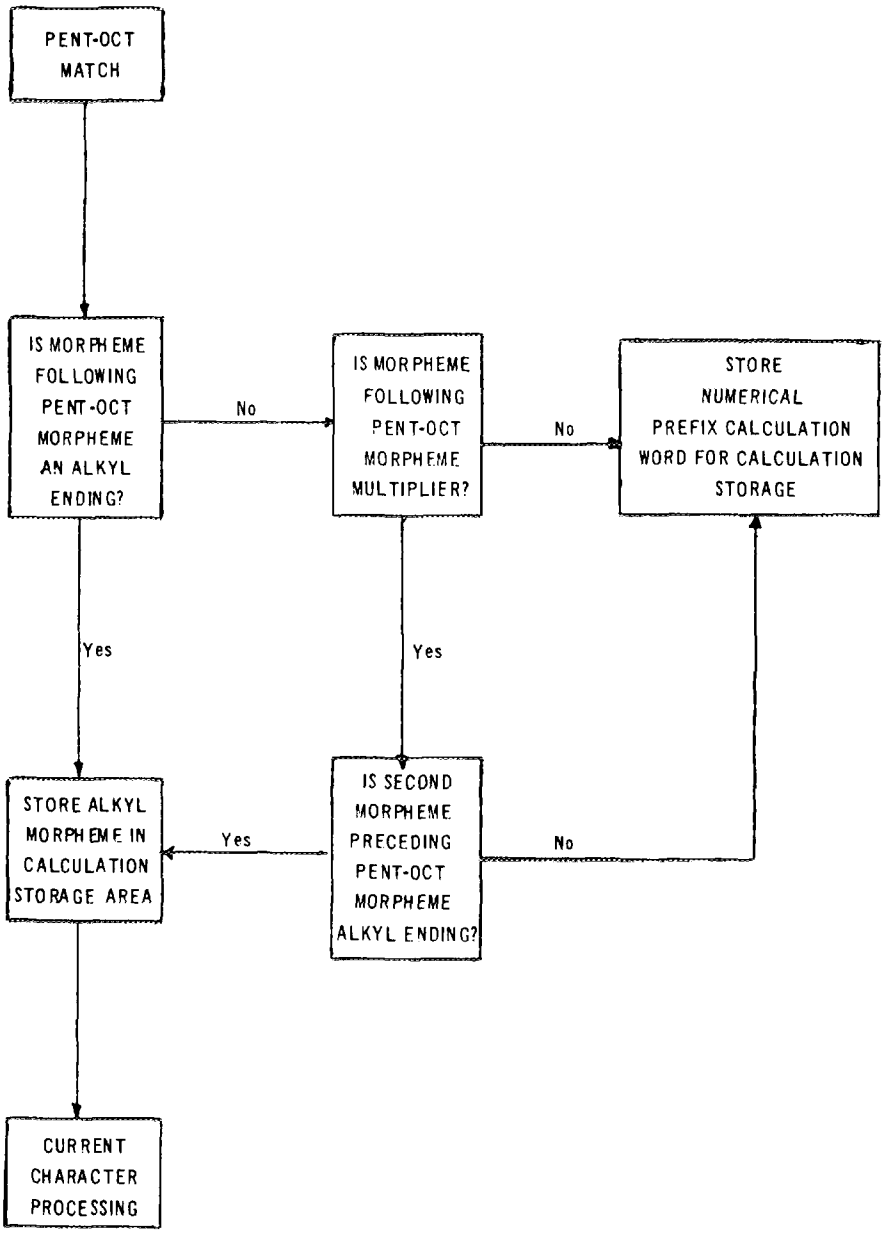


TABLE X. MOLECULAR FORMULA CALCULATION ROUTINE

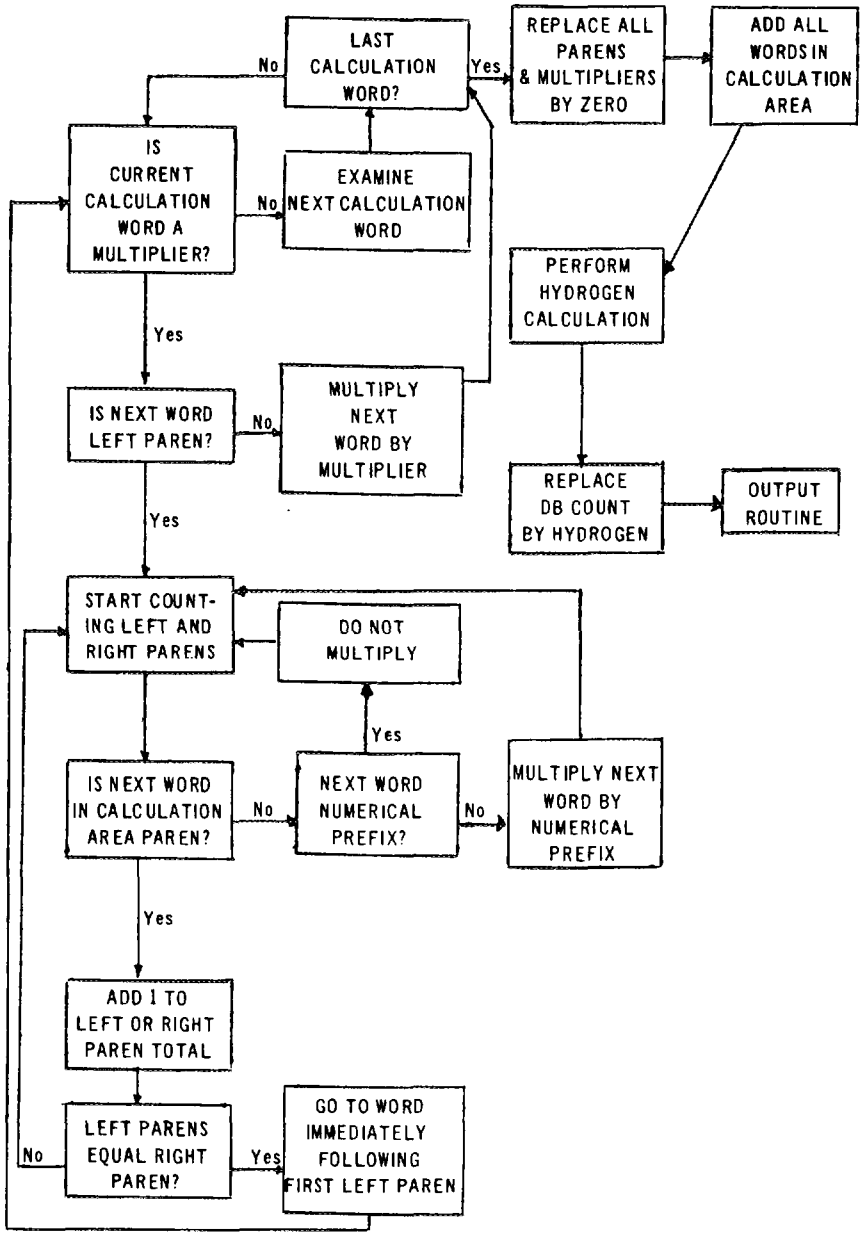


TABLE X

MOLECULAR FORMULA CALCULATION ROUTINE

1. Find a word which is a multiplier.
2. If the next word in calculation area is not a left paren, multiply it by the multiplier and continue looking for other multipliers.
3. If the next word is a left paren, starting keeping totals of left and right parens counting this as first left paren.
4. Examine each successive calculation word.
5. If it is not a paren, multiply it by the multiplier, unless it is a numerical prefix.
6. If it is a paren, add one to the left and right paren totals.
7. End process as soon as the left and right paren totals are equal.
8. Now go back to the word immediately following the first left paren and continue looking for multipliers. So process all multipliers in the calculation area.
9. Replace every paren word and every multiplier word in the calculation area by zero. Now add all words in the calculation area.
10. This gives preliminary total formula count. Now calculate H.
11. Replace DB value in preliminary total formula with the calculated H value.

The calculational value of each morpheme is stored as a twelve character number in which each successive pair of numbers represents iodine, double bonds, oxygen, nitrogen, sulfur and carbon. When the final calculation is made, the double bond position is replaced by the hydrogen count. Hence, nitro is stored as +0/01/02/01/00/00, iodo +1/00/00/00/00/00/ and methyl + 0/00/00/00/01. Since the Univac requires one character for sign, no formula containing more than nine iodine atoms can be tested in this equipment.

Sampling Method

The manual translation procedure was tested on dozens of chemicals. Some of these were deliberately selected as presenting difficulties. Others were randomly selected. For example, the deliberately chosen names included several that contained pent-oct ambiguous morphemes as e.g. *hexanitrohexadiene.* Others involved complex nesting of parens. The fifth example shown previously is typical of these.

Certain chemicals found in C.A. indexes did not calculate correctly for hydrogen. The morpheme *imino* was found to be used by C.A. quite inconsistently. A basic principle of good nomenclature is that structure and name should correspond. Names should not be based on the origins of compounds. A C.A. example is *1,1'-(ethylenediimino)di-2-propanol* which by I.U.P.A.C. nomenclature is *1,2-bis(2-hydroxy-propylamino)ethane. C.A.* violates the principle that each morpheme should consistently represent the same substituent. This name was omitted from the test as it would give wrong hydrogen count. In any computer program that would attempt to cover all systems, including C.A.'s, *imino* would require a special ambiguity-resolving routine.

When I was satisfied that I had deliberately tested all the morphemes in the dictionary a random sample of chemicals was obtained. This was done by asking a clerk to check the first chemical at the top of each column in the 1958 Subject Index to *Chemical Abstracts.* He was told to keep scanning until a name was located which could be obtained from the morphemes on the test list. This required the elimination of hundreds of chemicals which contain cyclic morphemes rather than acyclic. The following illustrate some of the samples located.

CA Page No.	*Molecular Formula*	*Chemical Name*
37	C_5H_9NO	2-hydroxy-2-methyl-butyronitrile
38	$C_5H_{10}N_2S$	2-amino-4-(methyl-thio)butyronitrile
39	$C_5H_{10}O_2$	4-methoxy-2-buten-1-ol
40	$C_5H_{11}NO_3$	methylnitro-2-butanol
52	$C_6H_9N_3$	3,3'-iminodipropionitrile
56	$C_6H_{11}NO_3$	6-amino-4-oxo-hexanoic acid
57	$C_6H_{12}N_2O_4$	2,3-dimethyl-2,3-dinitrobutane
58	$C_6H_{12}O_2S$	3-(propylthio)-propanoic acid
59	$C_6H_{13}NO$	4-dimethylamino-2-butanone
71	C_7H_9NO	3,4-dimethyl-2-oxopentenenitrile
73	$C_7H_{10}O_2$	3-ethylidene-2,4-pentanedione
80	$C_7H_{15}NO$	1-dimethylamino-2-methyl-3-buten-2-ol
80	$C_7H_{15}NO_3$	[bis(2-hydroxyethyl)amino]-2-propanone

I have intentionally listed the compounds for pages 37, 38 and 52 even though they do not come under the purview of this experiment. In spite of my instructions, it was apparently difficult for the person taking the sample to note that the *yro* and *ion* were not on the list of morphemes.

One other interesting example that had to be eliminated from the computer testing, but not the human testing, was the following: *N–[(2-[1,1-dimethyl-2-propynyloxy]ethoxy)methyl]diethylamine.* The use of the *N* as a locant was not anticipated in preparing the computer program. It would have to be added to the list of ignorable characters.

An additional random sample was taken from the *Merck Index.* This was done by taking a continuous series of chemicals in the cross-reference index. This gave quite a scattering of page numbers as is shown below:

Page	*M.F.*	*Chemical Name*
178	$C_4H_{11}N$	1-aminobutane
53	$C_4H_9NO_2$	4-aminobutanoic acid
738	$C_9H_{22}N_2$	2-amino-5-diethylaminopentane
1013	$C_2H_7NO_3S$	2-aminoethanesulfonic acid
315	C_2H_7NS	2-aminoethanethiol
4	$C_2H_6N_2$	α-amino-α-iminoethane
666	$C_5H_{11}NO_2S$	2-amino-4-methylthiobutanoic acid

Selections were still made that could not be handled by the experimental dictionary as e.g. sul*fonic* acid. Further, the use of *alpha(α)* as a locant was not anticipated for the computer program, though it could be easily added to the list of ignorable characters.

As a further test of the algorithm several chemists were asked to coin names that might be difficult to handle.

A few of these were 3,7-dimethyl-2,6-octadienal, 3,3′-dithiobis(2-aminopropanoic acid) and 1,4-bis(methanesulfonoxy)butane. The latter is not covered by the experimental dictionary.

Debugging

As a further test of the procedure, fifty of the randomly selected compounds were tested on the Univac I. The so-called debugging procedure uncovered dozens of coding mistakes in the computer program which had to be traced meticulously. Apparently the first twelve deliberately chosen compounds were well selected, as the computer went into loops on each one until the bugs were eliminated.

A More Significant Test

It is obviously important that the absolute validity of the algorithm be proven by more extensive sampling. However, the chemist knows intuitively, once he has used it, that it will work. When it fails, he will find ambiguities in the nomenclature as in the case of *imino*.

Of further importance was an informal test to verify the claim that chemists first draw a diagram in order to calculate molecular formulas. For this reason, I showed about a dozen chemists example *five* and asked each to calculate the formula. Invariably he would draw a diagram. All were surprised that the calculation could be reduced to such a brief algorithm. This confirms my belief that the algorithm can be an extremely useful teaching device. It certainly can be helpful to the indexer. Most graduate organic chemists already have memorized a large enough number of morphemes to calculate quite quickly without learning anything but the DB rules. This includes the more complex cyclic structures. Every steroid chemist knows that the steroid nucleus is C_{17} so it is quite simple for him to calculate steroid formulas no matter how complicated the name may be.
*See page 36.

The Bonding Morphemes

One particularly interesting product of this research has been the more precise definition of a class of so-called *endings* or suffixes for which, surprisingly enough, the chemist has no generic term. During the entire course of this investigation, difficulties were encountered in keeping programmers aware of the difference between an *alkyl* group and an *alkyl* or *alkane* ending. Neither of the latter two are accurate. Open chain hydrocarbons have the generic name *alkanes*. *Alkyl* is the generic term for hydrocarbon *radicals*. To use these terms to describe *alk-yl* suffixes is quite inaccurate. Furthermore, this does not associate all of the suffixes that can now be properly grouped in what I shall call the *bonding morphemes*. The members of this morpheme class are morphemes such as *ane*, *ene*, *yne*, *idene*, and *ium* since they contribute to the DB value of the chemical. In the pert-oct ambiguity-resolving routine, it would be more accurate to describe the operation in terms of bonding morphemes as the *alkyl* morphemes are really this group of bonding morphemes. It is interesting that to learn the algorithm completely from memory, the chemist need only learn the correct DB values for all morphemes, some of which may not be obvious. The chemist does not usually think of a triple bond as being two double bonds. Thus the DB value for *nitrilo*, *cyano*, *diazo*, and *yne* are the same i.e., DB_2.

Conclusions

I believe there are a number of important conclusions that can be drawn from this work. There can be no doubt that one can calculate molecular formulas from chemical nomenclature. The

grammatical work that remains to be completed is still quite large, but it does not appear to be so large that a group of chemists and linguists would have any difficulty completing it within a reasonable length of time. Further, if a computer is at their disposal, there are many shortcuts that could be taken in the analyses. If the grammatical work is expanded to include the type of syntactic analysis in which each morpheme is described as a part-of-speech, i.e. classified according to its membership in various grammatical categories, then it is quite possible to foresee a machine procedure which could generate standardized names. The same would be true of displaying structural diagrams. In fact, the latter problem is less sophisticated, in that there are a relatively small number of topological arrangements required in chemistry. The programming difficulties would arise in making the appropriate additions to the diagrams for substituent atoms. In the case of *nicotinoyl morpholine*, there is only one topological configuration, the hexagon, but the replacement of carbon by nitrogen and/or oxygen in the pyridine and morpholine rings requires considerable programming ingenuity. This work would be aided by the grammatical analyses.

It would also be safe to conclude that by similar procedures, one could analyze the chemical terminology of other languages and by establishing the transformations of that language, arrive at a method for translating chemical terminology quite easily. For certain languages, such as Russian, the work involved should not be very great as one can already, simply by transliteration of Russian nomenclature, understand most of the chemical names.

The linguistic approach to chemistry, i. e. chemico-linguistics holds great promise for chemist and linguist alike. For the chemist, it can mean greater precision in teaching and understanding nomenclature and even chemical classification per se. It is not improbable that a suitably written grammar of organic chemistry could help postulate new and interesting chemical structures. On the other hand, I believe that the field of chemistry offers the linguist a useful model for the study of normal discourse. If the problems of chemical nomenclature cannot be resolved by linguistic analysis, then I suspect that normal discourse will be much too formidable an obstacle. Certainly if we are to find methods of analyzing chemical texts for indexing and other purposes, we cannot expect better than a 50% resolution of the indexing problem in chemistry. More than 50% of the effort that goes into indexing chemistry is in the analysis of chemical names. A large part of the work that is done in reading chemical documents involves the recognition of dozens of chemical names, both new and old. We will have reaped a very poor harvest if we are able to describe the text of a chemical article grammatically without a corresponding ability to deal with the problem of synonymy.

TABLE XI

RANDOM SAMPLE OF CHEMICALS TESTED ON COMPUTER PROGRAM

butane = C_4H_{10}

2-aminoethanol = C_2H_7NO

1,4-bis(ethylamino)butane = $C_8H_{20}N_2$

1,3,5-heptatriene = C_7H_{10}

1,2,3,4,5,6,7-heptaiodooctane = $C_8H_{11}I_7$

2-[(3-aminopropyl)ethylamino]ethanol = $C_7H_{18}N_2O$

1,4-bis[bis(3-diethylaminopropyl)amino]butane = $C_{32}H_{72}N_6$

1-methylsulfonylbutane = $C_5H_{12}O_2S$

2-methylpropanedioic acid = $C_4H_6O_4$

1-propanethiol = C_3H_8S

3-pentanethione = $C_5H_{10}S$

1,6-dinitrohexane = $C_6H_{12}N_2O_4$

2,5-diaminohexanedioic acid = $C_6H_{12}N_2O_4$

4-oxo-heptanedioic acid = $C_7H_{10}O_5$

1-dimethylamino-2-methyl-3-buten-2-ol = $C_7H_{15}NO$

1-ethylamino-2-methyl-3-buten-2-ol = $C_7H_{15}NO$

2-(hydroxymethyl)-2-propyl-1,3-propanediol = $C_7H_{16}O_3$

3-ethyl-2-amino-3-pentanol = $C_7H_{17}NO$

8-hydroxy-6-octene-2,4-diynenitrile = C_8H_5NO

2-propenyl-2-pentenoic acid = $C_8H_{12}O_2$

2-ethylidene-3-methyl-1,5-pentanediol = $C_8H_{16}O_2$

2-nitro-2-pentyl-1,3-propanediol = $C_8H_{17}NO_4$

3-diethylamino-2-methyl-1-propanol = $C_8H_{19}NO$

5,5'-oxybis(2-methyl-2-pentanol) = $C_{12}H_{26}O_3$

1,1-diiodo-2-nitro-1-pentene = $C_5H_7I_2NO_2$

pentyl nitrate = $C_5H_{11}NO_3$

2,5-diiodo-hexanedinitrile = $C_6H_6I_2N_2$

1-aminobutane = $C_4H_{11}N$

4-aminobutanoic acid = $C_4H_9NO_2$

2-amino-1-butanol = $C_4H_{11}NO$

2-amino-5-diethylaminopentane = $C_9H_{22}N_2$

2-aminoethanethiol = C_2H_7NS

2-amino-5-hydroxypentanoic acid = $C_5H_{11}NO_3$

1-amino-1-iminoethane = $C_2H_6N_2$

2-amino-4-methylthiobutanoic acid = $C_5H_{11}NO_2S$

3-methyl-1-pentyn-3-ol = $C_6H_{10}O$

1,3-butadiene = C_4H_6

bis(hydroxyethyl)amine = $C_4H_{11}NO_2$

2,2-bis(hydroxymethyl)-1,3-propanediol = $C_5H_{12}O_4$

TABLE XI (cont.)

2-ethoxyethanol = $C_4H_{10}O_2$
dimethylenimine = C_2H_3N
3,7-dimethyl-2,6-octadienal = $C_{10}H_{16}O$
3,3'-dithiobis-(2-aminopropanoic acid) = $C_6H_{12}N_2O_4S_2$
1-iodo-3-iodomethyl-5-methylheptane = $C_9H_{18}I_2$
1,4-diiodo-2-(methylbutyl)-butane = $C_9H_{18}I_2$
methylsulfonylethane = $C_3H_8O_2S$
(2-hydroxyethyl)-4-hydroxymethyl)-3-propyl-1,6-hexanediol = $C_{12}H_{26}O_4$
methylthiopropane = $C_4H_{10}S$
1-(propylsulfinyl)butane = $C_7H_{16}OS$
ethylsulfinylethane = $C_4H_{10}OS$
ethanamide = C_2H_5NO
butanediamide = $C_4H_8N_2O_2$
methylthiopropane = $C_4H_{10}S$
nitrosobutane = C_4H_9NO
ethylmethyl peroxide = $C_3H_8O_2$
iodosoethane = C_2H_5IO
iodoxypropane = $C_3H_7IO_2$
sulfopropanoic acid = $C_3H_6O_5S$
ethanethial = C_2H_4S
trichloromethane = $CHCl_3$
tetranitromethane = CN_4O_8
1-nitro-1,1,2,2,2-pentachloroethane = $C_2Cl_5NO_2$
hexachloroethane = C_2Cl_6
1,1,2-trichloroethane = $C_2H_3Cl_3$
octachloropropane = C_3Cl_8
propylnitrate = $C_3H_7NO_3$
1,1,1,3,3-pentachloro-2,3-dinitro-2-trichloro-methylpropane = $C_4Cl_8N_2O_4$
4-chloro-3-butyn-1-ol = C_4H_5ClO
2-methyl-1,2-dinitropropane = $C_4H_8N_2O_4$
1,4-diamino-2-butanone = $C_4H_{10}N_2O$
1,3,3,4,4-pentachloro-2-methylcyclobutene = $C_5H_3Cl_5$
penten-4-ynol = C_5H_6O
4,5,5-trichloro-4-pentenylamine = $C_5H_8Cl_3N$
dimethylcyclopropane = C_5H_{10}
chloropentanol = $C_5H_{11}ClO$
pentachlorobenzene = C_6HCl_5
2-aminochloronitrophenol = $C_6H_5ClN_2O_3$
benzenediol = $C_6H_6O_2$
2,6-dichlorocyclohexanone = $C_6H_8Cl_2O$
1,1,1-trichloromethyl-3-penten-2-ol = $C_6H_9Cl_3O$
1-cyclopentene-1-methanol = $C_6H_{10}O$
chlorocyclohexane = $C_6H_{11}Cl$
2-amino-4-butyl-6-nitrophenol = $C_{10}H_{14}N_2O_3$
(1-cyclohexen-1-yl) butanone = $C_{10}H_{16}O$
2-phenyl-2,4,6-cycloheptatrien-1-one = $C_{13}H_{10}O$
7-(2,4,5-trichlorophenoxy)heptanoic acid = $C_{13}H_{15}Cl_3O_3$
ethyl 2-cyano-5-phenyl-2,4-pentadienoate = $C_{14}H_{13}NO_2$
7-(4-dimethylaminophenyl)-2,4,6-heptatrienenitrile = $C_{15}H_{16}N_2$
1-3-bis(aminophenoxy)-2-propanol = $C_{15}H_{18}N_2O_3$
4,6-dibutyl-3-methyl-2,4-dinitro-2,5-cyclohexadien-1-one = $C_{15}H_{22}N_2O_5$
2,4-dimethyl-3-octyl-2-cyclopenten-1-one = $C_{15}H_{26}O$
2-nitro-4-phenyl-1-naphthol = $C_{16}H_{11}NO_3$
1-(nitrophenyl)-4-phenyl-2-butene-1,4-dione = $C_{16}H_{11}NO_4$
2-(naphthyl)-2-cyclohexen-1-one = $C_{16}H_{14}O$
diphenyl-3-butynol = $C_{16}H_{14}O$

APPENDIX

I.U.P.A.C. Organic Chemical Nomenclature
A Summary of Principles Including a Detailed Example of its use both in Recognition and Generation of Systematic Names

In summarizing the basic principles of I.U.P.A.C. organic nomenclature for the non-chemist emphasis has been placed on didactive explanations that will help in the recognition of the meaning of chemical names, rather than complete rules for the generation of names. The latter would require a knowledge of chemistry at least to the extent of understanding structural diagrams. This is not even necessary for the acyclic straight chain hydrocarbons covered in this experiment. Therefore, by following the instructions for naming hydrocarbon derivatives, a non-chemist should have no difficulty creating perfectly reasonable and accurate names for simple chemicals. For the more complex molecules, I suspect he would have no more and possibly less difficulty than the chemist who comes to the subject with certain preferences based on his knowledge of chemistry.

Punctuation

Commas are used between numerals which refer to identical operations as in *1,2,3-tribromohexane.*

Colons are used between groups of numerals for similar but distinct operations as in *1,2:5,6-diisopropylidenesorbitol.*

Numerals should be placed immediately in front of the syllables to which they refer as e.g. *2-bromohexane* rather than *bromo-2-hexane; hexan-2-ol* rather than *hexanol-2.* However, in the U.S. 2-hexanol would be rather commonly encountered. The numeral designates the number of the carbon atom in the longest chain of carbon atoms contained in the chemical. The variations in the use of numerals are legion and present a major obstacle to comprehension, especially in French and German literature. In some systems Greek letters are used instead of numerals. Amino acids are popularly numbered this way as in *β-hydroxyalanine*, which is also *2-amino-3-hydroxypropanoic acid* also known as *serine.*

Order of Substituents

Prefixes are arranged in *alphabetical order.* The atoms and groups are alphabetized first and the multiplying prefixes are then inserted as in: *2-bromo-1-chloro-hexane; 4-ethyl-3-methyl-hexane;* and *1,1,1-trifluoro-3,3-dimethylpentane.*

Elision

The terminal *e* is elided before a vowel of an organic suffix, but not in cases where the following letter is a consonant. *Propane* becomes *propanone; hexan-2-one* becomes *hexane-2,3-dione.*

Hyphens

These are used between two identical letters to avoid ambiguity as in tetra-amino. The Chemical Society uses hyphens also when partial names end in a voiced vowel or y as e.g. in *amino-*derivative, *thia-*compound, *methoxy-*group, but not after a consonant in such places as methyl derivative, amide group. In English, chemical words do not end in vowels.

Parentheses

Parens are used when necessary to clarify the limits of operations but not unnecessarily. If a string of morphemes is contained in parens which is preceded by a numeral, this means that the entire parenthesized expression is a substituent of a parent structure. For example, *1-(4-amino-2-ethylphenyl)-butanol* means that the entire expression *4-amino-2-ethylphenyl* is attached to the first atom in a four carbon *(but)* chain. The word *mono* is understood but rarely used. However, if the chemical were *1,2 bis-(4-amino-2-ethylphenyl) butanol* the entire parenthesized expression would be multiplied by two, i.e. it occurs at both the first and second carbon atoms in the chain C–C–C–C.

Terminology

Parent is a very ambiguous term in chemical nomenclature, especially when one considers the rules for deciding which morpheme in a name shall be considered the parent morpheme. However, no matter what name is chosen the *parent* morpheme refers to that group of atoms to which all other groups of atoms in the molecule are attached. Thus *benzene* is the parent in *nitrobenzene* and *ethane* is the parent in *ethanol.* This term no longer has any chemical significance which, at one time, was true when chemicals were named on the basis of the shortest chain length.

Group or radical. Any group of atoms commonly occurring together is called a group or radical. Most of these are single morphemes but some are pairs of morphemes. CH_3 is a *methyl* group consisting of the morphemes *meth* and *yl*. However, OH is the hydroxy group.

Function or Functional Group

A *functional group* is a group of atoms which defines the mode of activity of a chemical. The hydroxy group gives alcoholic properties to an alcohol. A ketone owes its properties to the oxygen atom which is doubly bonded to carbon. The distinction between what is functional and what is not is frequently difficult to make, but is an important artefact in naming chemicals regardless of how they act.

Types of Names

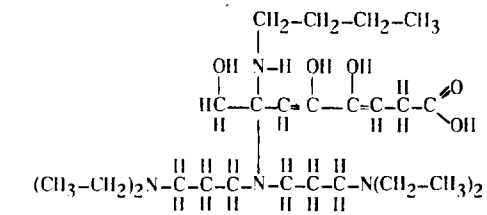
There are several types of names encountered in systematic nomenclature aside from the previously discussed *trivial* and semi-systematic names. There are names which involve *substitution*, where one hydrogen atom is replaced by a group or another element, as in *pentanol*, where one hydrogen atom is replaced by the hydroxy radical or group. There are *replacement* names, where one atom such as sulfur replaces another, such as oxygen, as for example *propanol* and *propanethiol*, which are respectively C–C–C–OH and C–C–C–SH.

A *subtractive* name involves the removal of specified atoms as e.g. in aliphatic names ending in *ene* or *yne* exemplified by *hexene* or *hexyne* where hydrogen atoms are removed by the creation of double or triple bonds between carbon atoms –– C–C–C–C–C=C.

There are other types of names such as *radicofunctional*, a name formed from a radical and functional class name such as *ethyl alcohol; additive* names such as *styrene oxide*, conjunctive names such as *naphthaleneacetic acid*, and fusion names such as *benzofuran* and other cyclics. However, in this brief survey, we will be primarily concerned with systematic names, i.e. names "composed wholly of specially coined or selected syllables, with or without numerical prefixes" [cf. I.U.P.A.C.: *Nomenclature of Organic Chemistry*. London: Butterworths, 1958(p. 4)].

What's In a Name?

When the layman sees a chemical name like *7-bis(3-diethylaminopropyl)amino-7′butylamino-4,5,8-trihydroxyoct-3,5-dienoic acid*, he probably wonders how it is possible for chemists to make sense out of it. The structural diagram for this chemical is



However, chemical names are surprisingly simple to understand and a large number of those made can be derived from a relatively short list of morphemes, such as that which was used in my experiments (see Table VI).

Principal Functional Group

The first thing that must be done in understanding, or for that matter in creating a chemical name is to "*seek out the functional groups.*" (Cahn, opus cited p. 43.) The *senior*, i.e. *principal functional group* sets the whole pattern of nomenclature and numbering. Unfortunately this is not always as simple as it sounds, though in the example above it is quite simple. It is worth noting that among others Degering (cf. *Organic Chemistry -- An Outline of the Beginning Course Including Material for Advanced Study*; 6th Ed. New York: Barnes & Noble, 1957) completely avoids a discussion of this problem of naming so-called *complex* functions, that is, chemicals with more than one functional group. He is well advised to do so because there is no rational way of explaining this principle though Chemical Abstracts and others will specify a preferred order of precedence--acid before aldehyde, aldehyde before ketone, ketone before alcohol, etc. Cahn would agree with this order I.U.P.A.C. does not stipulate a preferred order. Since most chemicals in the U.S. and Great Britain are named by this order one can conclude, in the example shown, that the principal function is the acid function. It is assumed by now that the reader understands that each chemical name can be parsed quite simply into a series of short letter sequences, i.e. morphemes. By reference to Table XII, it will be noted that each of these morphemes has an associated meaning. The *oic acid* at the end of this name is such a morpheme as are *di* and *en* which precede it. *En* is a bonding morpheme, that is, it denotes *unsaturation* in the basic carbon chain of the molecule. Unsaturation refers to the removal of hydrogen atoms attached to carbon atoms to form double bonds. The entire structure of organic nomenclature is based on the theory of covalent bonds.

Most Unsaturated Straight Chain

In naming this chemical no difficulty would arise concerning the next principal group as

there is no choice here between two sometimes perplexing alternatives of a shorter chain with greater unsaturation and a longer chain with no or less unsaturation. If there were, then the chain with the most double and triple bonds would be selected. This would be the case, e.g. if there were a side chain containing two additional double bonds. As the second priority item in naming a chemical, the saturation is indicated second from the right. In other words, the so-called principal functions come at the end of the name preceded by bonding morphemes when this is possible.

The Longest Chain

The third criterion for selecting the proper name is the principle of the longest chain. By this is meant not the longest chain of atoms, but the longest chain of consecutive carbon atoms. There is, indeed, a school of thought that prefers the principle whereby the longest chain is used, regardless of the atoms involved. A good case can be made for it in many instances. In this particular chemical, the longest chain of carbon and nitrogen atoms is fourteen. The longest carbon chain is eight atoms long and that is why the next morpheme to the left of *dien* is *octa* signifying an eight carbon chain (C–C–C–C–C–C–C–C).

Numbering

After making the decision as to which sequence of atoms in the molecule will become the *parent, then one numbers each of the contiguous atoms giving the atom to which the functional group* is attached the lowest number. In our example, the *oic acid* function is the principal function, consequently the numbering pattern will be (HO)O=C–C–C=C–C=C–C–C. This will explain the numerals preceding diene as the two double bonds are located between carbon atoms 3 and 4 and atoms 5 and 6.

Substituents or Prefixes

Once the selection of the parent chain has been completed, as well as adding as suffixes, the bonding morphemes and the principal functions, it only remains to name the substituents or side chains, all of which may be regarded as radicals, groups, or sub-names depending upon the complexity of the chemical. In this particular case there are three hydroxy groups at the third, fourth, and *eighth atoms*. They are specified by using the numerals 3,4,8 followed by the numerical prefix *tri* followed in turn by the morpheme *hydroxy*, hence *trihydroxy*. The remaining substituents in this name are themselves substituted as e.g. *butylamino* which means that there is a *nitrogen* atom attached to the seventh atom in the *octane* parent structure. Ordinarily, amino implies the replacement of one hydrogen atom by the amino group (NH_2), but in this case, one of the amino hydrogens is also replaced by a radical, the *butyl* radical, which is composed of a four carbon chain. Hence, *butylamino* is $CH_3CH_2CH_2CH_2NH-$. By a similar building up process, the last portion of

this name, *(diethylaminopropyl) amino* is the following: $(C_2H_5)_2$–N–CH_2–CH_2–CH_2–N–. However, since the parenthesized expression is preceded by *bis*, it simply means that the other bond on the right most nitrogen has the same chain repeated, which means we really have, for this side chain $[(C_2H_5)_2$–N–CH_2–CH_2–CH_2–$)_2$N–, i.e. *bis(diethylaminopropyl) amino*. The 3- preceding diethyl simply specifies that the left most amino group is attached to the third carbon atom in the *propyl* chain.

This sketch of the rules and explanation of this very complex example does not cover all of the problems. Of interest to the linguist is the choice of allomorph to be made e.g. for *OH*, the hydroxy group rather than *ol*. It is only when the principal function is an alcohol that this latter morpheme is used. Were the carboxyl group (oic acid) to be replaced by another hydroxy group, the name of this chemical would change considerably, but primarily by the elimination of the prefix *3,4,8-trihydroxy* and the addition of *tetrol* as a suffix giving us a name ending in *octa-3,5-dien-1, 4,5,8-tetrol*.

Since chemicals can be prepared with a multitude of different permutations and combinations, the reader can well imagine the difficulties one may encounter when having to make a preferred choice. It is no small wonder that chemists arrive at different names. If considerations of cyclic nomenclature are introduced, then the absurdities of nomenclatural logic increase to the point where there is mass confusion. If the principal function is attached to a ring, i.e. cyclic, then it is the cyclic system which is given priority over the acyclic chain, no matter how long, but if the principal functional group is attached to a chain and that in turn to a ring, the British would treat the cyclic radical as a substituent, while the C.A. indexer would take into consideration the complexity of the cyclic substituent and more than likely call it the principal function.

In closing this discussion, it is worth emphasizing that in spite of the variations in naming chemicals, one generally will have no difficulty in figuring out the chemical involved, because it can always be pieced together by reference to the dictionary of morphemes. If that were not the case communication between chemists would have ceased long ago. This is not to underestimate the difficulties of decipherment. In general such difficulties arise from the fact that the distraught chemist trying to use "systematic" nomenclature, invariably forgets one of the rules and in his confused state generates an ambiguous name. He does not always take the trouble to ask another chemist to try deciphering the name he has chosen. Wiser chemists rely strictly on structural diagrams. Perhaps this accounts for the success of the Japanese chemists who are used to working with ideographs. In this connection, a closing quotation from the British Chemical Society's heated disucssion of the Geneva Conference in which it is said "Prof. P.F. Frankland thought names unnecessary, and that it would be better for the purposes of a register to use formulae." (Armstrong, H.E., opus cited p. 130) seems both pertinent and ironically, prophetic.

Table XII can also be used as a condensed review of I.U.P.A.C. nomenclature. It covers twenty-three primary generic groups of chemicals synthesized by the organic chemist. Each type

is shown by indicating an *R* group, the conventional symbol for *radical* attached to the appropriate functional group. Following the generic name, the most commonly used morpheme is listed. For any specified value of R and/or R′, one can quickly determine the sort of chemical name to expect. In this experiment, particular attention was given to compounds where the R values would consist of the homologous series *meth*, *eth*, prop, *but*, *pent*, *hex*, *hept*, and *oct*, i.e. where R equals one, two, three, etc. carbon atoms. Finally, the calculational value for each morpheme is shown. This can be used in applying the algorithm for the calculation of molecular formulas. A more complete list of the morphemes used in the experiment is shown in Table VI on pages 30–31.

TABLE XII. SUMMARY OF I.U.P.A.C. NOMENCLATURE

Structure	Generic Name	Morpheme	Value
R–CH_3	alkanes	ane	DB_0
R=CH_2	alkenes	ene	DB_1
R≡CH	alkynes	yne	DB_2
R–OH	alcohols	ol	O_1
R–SH	mercaptans	thiol	S_1
R–	radicals	yl	(+)
R–O–R′	ethers	oxy	O_1
R–S–R′	sulfides	thio	S_1
R–SO–R′	sulfoxides	sulfinyl	S_1+O_1
R–SO_2–R′	sulfones	sulfonyl	S_1+O_2
R–CH=O	aldehydes	al	O_1+DB_1
R–CH=S	thioaldehydes	thial	S_1+DB_1
R–C(R′)=O	ketones	one	O_1+DB_1
R–C(R′)=S	thioketones	thione	S_1+DB_1
R–COOH	carboxylic acids	oic acid	O_2+DB_1
RCSOH	thio acids	thioic acid	$S_1+O_1+DB_1$
RCOOR′	salts & esters	oate	O_2+DB_1
R–COX	acid halides	oyl halide	$O_1+DB_1+X_1$
$RCONH_2$	amides	amide	$O_1+DB_1+N_1$
R–CN	nitriles	nitrile	DB_2+N_1
R–NO_2	nitro derivatives	nitro	$O_2+DB_1+N_1$
R–NO	nitroso	nitroso	$O_1+DB_1+N_1$
$RONO_2$	nitrates	nitrate	$O_3+DB_1+N_1$

BIBLIOGRAPHY

Armstrong, H. E.: Contributions to an International System of Nomenclature. The Nomenclature of Cycloids. *Proc. Chem. Soc.*, 1892,127.

Bloomfield, L.: *Language*. New York: Holt & Co., 1933.

CA: Naming & Indexing of Chemical Compounds by Chemical Abstracts. Columbus: Chemical Abstracts, 1957.

Cahn, R. S.: *An Introduction to Chemical Nomenclature*. London: Butterworths, 1959.

Cahn, R. S. & Cross, L. C.: *Handbook for Chemical Society Authors*. London: The Chemical Society, 1960.

CLR: *Fourth Annual Report, Council on Library Resources*. Washington: The Council, 1961.

Crane, E.J.: *CA Today – The Production of Chemical Abstracts*. Washington: Amer. Chem. Soc., 1958.

Dyson, G.M.: *A New Notation and Enumeration System for Organic Compounds*. New York: Longmans, 1949.

Frome, J.: *Semi-Automatic Indexing and Encoding. Research and Development Report No. 17*. Washington: U. S. Patent Office, 1959.

Garfield, E.: Communication concerning the use of machines in facilitating documentation. *Chem. Eng. News*, 30:5232,1952.

Garfield, E.: *Preparation of Printed Indexes by Automatic Punched-Card Equipment – A Manual of Procedures*. Baltimore: Johns Hopkins University Medical Indexing Project, 1953.

Garfield, E.: Preliminary Report on the Mechanical Analysis of Information by use of the 101 Statistical Punched-Card Machine. *Am. Documentation*, 5:7, 1954.

Garfield, E.: Preparation of Subject Heading Lists by Automatic Punched-Card Techniques, *J. Documentation*, 10:1, 1954.

Garfield, E.: Forms for Literature Citations, *Science*, 120:1039,1954.

Garfield, E.: Preparation of Printed Indexes by Machines, *Am. Documentation*, 6:68,1955.

Garfield, E.: Citation Indexes for Science, *Science*, 122:108,1955.

Garfield, E.: Breaking the Subject Index Barrier, *J. Pat. Off. Soc.*, 39:583, 1957.

Garfield, E.: A Unified Index to Science, *Proc. Intl. Conf. on Scientific Information, Vol. 1*. Washington: National Academy of Sciences, 1959.

Garfield, E.: The Steroid Literature Coding Project, *Chem. Literature*, 12(3):6,1960.

Garfield, E.: Index Chemicus Molecular Formula Index, *Index Chemicus*, First Cumulative Index Issue:1961,33.

Harris, Z. S.: *Methods in Structural Linguistics*. Chicago: Univ. of Chicago Press, 1951.

Harris, Z.S.: Linguistic Transformations for Information Retrieval, *Proc. Intl. Conf. on Scientific Information, Vol. 2*. Washington: National Academy of Sciences, 1959.

BIBLIOGRAPHY (continued)

Harris, Z.S.: Hiz, H., Joshi, A. K., Kaufman, B., Chomsky, C., and Gleitman, L.: Transformations & Discourse Analysis. *Annual Report of the Computing Center.* Philadelphia: Univ. of Pennsylvania, 1960.

Harris, Z.S.: Hiz, H., et al. *Transformations and Discourse Analysis Projects.* Philadelphia: Univ. of Pennsylvania, Department of Linguistics, 1959–61.

Himwich, W. A., Field, H., Garfield, E., Whittock, J. and Larkey, S. V.: *Welch Medical Library Indexing Project Final Reports.* Baltimore: Johns Hopkins Univ., 1951, 1953, 1955.

I.U.P.A.C.: Definitive Rules for Nomenclature of Organic Chemistry, *J. Am. Chem. Soc.*, 82:5545,1960.

Opler, A., and Baird, N.: Display of Chemical Structural Formulas as Digital Computer Output, *Am. Documentation*, 10:59,1958.

Patterson, A.M.: Definitive Report of the Commission on the Reform of the Nomenclature of Organic Chemistry, *J. Am. Chem. Soc.*, 55:3905,1933.

Patterson, A. M.: *Words about Words.* Washington: American Chemical Society, 1957.

Patterson, A. M., Capell, L. T. and Walker, D. F.: *The Ring Index – A List of Ring Systems used in Organic Chemistry.* 2nd Edition. Washington: American Chemical Society, 1960.

Pictet, A.: Le Congres International de Geneve pour la Reforme de la Nomenclature Chimique, *Arch. Sci. Phys. Nat.*, 27:485,1892.

Rabinow, J.: *Character Recognition Machines.* Washington: Rabinow Engineering Co., 1961.

Soffer, M. D.: The Molecular Formula generalized in terms of cyclic elements of structure. *Science*, 127:880,1958.

Stock, C.C.: *A Method of Coding Chemicals for Correlation and Classification.* Washington: National Academy of Sciences, 1950.

Terentiev, A. P., Kost, A. N., Tsukerman, A. M. and Potapov, V. M.: *Nomenklatura Organicheskikh Soedmionii.* Moscow: Akademiya Nauk SSSR, 1955.

Tiemann, F.: Ueber die Beschlusse des internationalen, in Genf vom 19 bis 22. April 1892 Versammelten Congresses zur Regelung der chemischen nomenclatur, *Ber. d. Deut. Chem. Gesell*, 26:1595,1892.

Tsukerman, A. M. & Terentiev, A. P.: Chemical Nomenclature Translation. *Proc. Intl. Conf. for Standards on a Common Language for Machine Searching & Translation.* Vol. I. New York: Interscience Press, 1961.

Waldo, W.H. and de Backer, M.: Printing Chemical Structures Electronically. *Proc. Intl. Conf. on Scientific Information. Vol. II.* Washington: National Academy of Sciences, 1959.

Wiswesser, W. J.: *A Line Formula Chemical Notation.* New York: Thos. Crowell, 1954.