(20) Sheridan, R. P.; Nilakantan, R.; Rusinko, A.; Bauman, N.; Haraki, K. S.; Venkataraghavan, R. 3DSEARCH: a System for Three-Dimensional Substructure Searching. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 255-260.
(21) Mitchell, E. M.; Artymiuk, P. J.; Rice, D. W.; Willett, P. Use of Techniques Derived from Graph Theory to Compare Secondary Structure Motifs in Proteins. *J. Mol. Biol.* **1990**, *212*, 151-166.
(22) von Scholley, A. A Relaxation Algorithm for Generic Chemical Structure Screening. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 235-241.
(23) Flanders, P. M.; Hunt, D. J.; Reddaway, S. F.; Parkinson, D. Efficient High Speed Computing with the Distributed Array Processor. In *High Speed Computers and Algorithm Organization*; Kuck, D. J., Lawrie, D. H., Sameh, A., Eds.; Academic Press: New York, 1977.
(24) Hunt, D. J.; Reddaway, S. F. Distributed Processing Power in Memory. In *The Fifth Generation Computer Project*; Pergamon Infotech: London, 1983.
(25) Parkinson, D. The Distributed Array Processor (DAP). *Comput. Phys. Commun.* **1983**, *28*, 325-336.
(26) Parkinson, D.; Litt, J., Eds. *Massively Parallel Computing with the DAP*; Pitman: London, 1990.
(27) Willett, P.; Rasmussen, E. M. *Parallel Database Processing. Text Retrieval and Cluster Analysis Using the Distributed Array Processor*; Pitman: London, 1990.
(28) Unger, S. H. A Computer Oriented towards Spatial Problems. *Proc. Inst. Radio Eng. (USA)* **1958**, *46*, 1744-1750.
(29) Slotnick, D. L.; Borck, W. C.; McReynolds, R. C. The SOLOMON computer. *AFIPS Conf. Proc.* **1962**, *22*, 97-107.
(30) Potter, J. L., Ed. *The Massively Parallel Processor*; MIT Press: Cambridge, MA, 1985.
(31) Hillis, W. D. The Connection Machine. *Sci. Am.* **1987**, *256* (6), 86-93.
(32) Tucker, L. W.; Robertson, G. G. Architecture and Applications of the Connection Machine. *Computer* **1988**, *21* (8), 26-38.
(33) Reddaway, S. F. Achieving high performance applications on the DAP. In Jesshope, C. R. *CONPAR 88*; Jesshope, C. R., Reinartz, K. D., Eds.; Cambridge University Press: Cambridge, 1989.
(34) Moitra, A.; Iyengar, S. S. Parallel Algorithms for some Computational Problems. *Adv. Comput.* **1987**, *26*, 93-153.
(35) Cringean, J. K.; Manson, G. A.; Willett, P.; Wilson, G. A. Efficiency of Text Scanning in Bibliographic Databases Using Microprocessor-Based, Multiprocessor Networks. *J. Inf. Sci.* **1988**, *14*, 335-345.
(36) Wilson, T. Ph.D. Thesis; University of Sheffield (in preparation).
(37) van't Riet, B.; Kier, L. B.; Elford, H. L. Structure Activity Relationships of Benzohydroxamic Acid Inhibitors of Ribonucleotide Reductase. *J. Pharm. Sci.* **1980**, *69*, 856-857.
(38) Hansch, C.; Yoshimoto, M. Structure-Activity Relationships in Immunochemistry. 2. Inhibition of Complement by Benzamidines. *J. Med. Chem.* **1974**, *17*, 1160-1167.
(39) Richard, A. J.; Kier, L. B. Structure-Activity Analysis of Hydrazide Monoamine Oxidase Inhibitors Using Molecular Connectivity. *J. Pharm. Sci.* **1980**, *69*, 124-126.
(40) Adamson, G. W.; Bawden, D. A Method of Structure-Activity Correlation Using Wiswesser Line Notation. *J. Chem. Inf. Comput. Sci.* **1975**, *15*, 215-220.
(41) Chen, B. K.; Horvath, C.; Bertino, J. R. Multivariate Analysis and Quantitative Structure-Activity Relationships. Inhibition of Dihydrofolate Reductase and Thymidylate Synthetase by Quinazolines. *J. Med. Chem.* **1979**, *22*, 483-491.
(42) Parkinson, D.; Liddell, H. M. The Measurement of Performance on a Highly Parallel System. *IEEE Trans. Comput.* **1983**, *C32*, 32-37.

Computer Storage and Retrieval of Generic Chemical Structures in Patents. 11. Theoretical Aspects of the Use of Structure Languages in a Retrieval System

WINFRIED DETHLEFSEN
BASF, Ludwigshafen/Rhein, Germany

MICHAEL F. LYNCH,* VALERIE J. GILLET, GEOFFREY M. DOWNS, and JOHN D. HOLLIDAY
Department of Information Studies, University of Sheffield, Sheffield S10 2TN, England

JOHN M. BARNARD
Barnard Chemical Information Ltd., 46 Uppergate Road, Stannington, Sheffield S6 6BX, England

Received November 16, 1990

A rational basis for discussion of issues relating to the storage and retrieval of generic chemical structures is developed in this paper and those which follow. It rests on well-known logical and linguistic foundations, and seeks to establish a consistent conceptual framework for considering generic structures as they occur in patents and as represented for storage and retrieval in information systems. The syntax, semantics, and pragmatics of chemical structure languages, in general, are described, together with the meaning-relations between the notation, the intension, and the extension of a structural expression. Development of this basis provides a framework for considering issues of the representation of generic structures in formally defined languages, such as GENSAL, together with the process of translation from chemists' language into GENSAL, the surface language, and of further translation into other internal representations, including the ECTR (Extended Connection Table Representation), and ring, fragment, and reduced graph screens for processing and searching. The question of the definiteness of structure representations and its consequences for searching are discussed, together with formal properties of structural expressions in the GENSAL system, and the applicability of a variety of algorithms. Finally, the relations between query and file structure languages are described.

1. INTRODUCTION

The representation of generic structures for retrieval poses complex questions, many of which have already been the subject of publications from Sheffield, e.g., by Downs et al.[1] and preceding articles in this series, and elsewhere, in particular by Shenton et al.[2,3] and by Fisanick.[4,5]

The importance of the area for chemical industry is reflected in the introduction of the Markush DARC system in 1989 by Derwent Publications Ltd., with Questel S.A. and INPI (the French Patent Office), and of the MARPAT system in 1990 by Chemical Abstracts Service. International Documentation in Chemistry GmbH (IDC) is also active in innovation, creating a database of generic structures in GENSAL in order to generate GREMAS fragment codes automatically.[6]

GENSAL, first described by Barnard et al.,[7] is an artificial and formally defined language, the purpose of which is to record chemical structure information in patents in such a way that the resultant expressions are amenable to algorithmic manipulation in order to support a variety of types of structure-based searches. Its use involves translations from the language used by chemists, patent agents, and lawyers in patent documents into the formally defined language. GEN-

0095-2338/91/1631-0233$02.50/0

SAL itself, a 'surface' language, is designed to achieve as high a degree of representational fidelity and similarity to chemists' language as possible. In this regard, it is a formalization of chemists' language. Expressions in GENSAL are in turn translated automatically by the GENSAL interpreter into an internal representation, the ECTR (Extended Connection Table Representation),[8,9] and into other forms, for searching purposes.

A sample database was created in 1983–1984[10] to test the viability of the language for database creation purposes; it used a version of the Feldman teletype-compatible structure-drawing system. More recently, IDC has commissioned a greatly improved graphics-based database creation module, Microgensip;[6] the Markush DARC and MARPAT systems also involve database creation modules.

Discussions of the requirements of the GENSAL language and experience in its use, as well as insights gained into the subsequent algorithmic operations on GENSAL records in Sheffield and at IDC, have raised many fundamental issues to which little thought had earlier been given. These issues have not yet received the attention which is their due, in particular, questions of the relationship between information on generic structures as it is recorded in patent documents and in the GENSAL language and in the representations derived from GENSAL for search purposes. The issues include the process of translation from the terms in which structural (and related) information is provided in patents into the formal system representation, its manipulation within the system, the specification of structural queries by users in as flexible and responsive a manner as possible, and the methods by which the matches are determined. Related issues have earlier been considered by Gordon.[11]

This paper and those which follow it thus seek to lay a rational foundation for consideration of these relations and of issues arising from them. They stand together with subsequent papers, in which the practical consequences of the theory for implementation as algorithms are described.

The readability of the remainder of this paper may be made easier by giving a survey in the form of a table of contents with additional annotations in which terms (introduced or explained in the respective section) are set in boldface type:

2. SYNTAX, SEMANTICS, AND PRAGMATICS OF STRUCTURE LANGUAGES

Chemical structures are stored and retrieved by means of signs. Signs used in a language are related in three ways: to other signs in the language, to the objects designated, and to the users of the language. Thus, according to the model of semantics proposed by Morris,[12] the use of signs has three dimensions: the syntactic, the semantic, and the pragmatic.

This distinction between syntactic, semantic, and pragmatic aspects of a language used in the linguistics of natural and of formal languages may also be applied to those languages whose objects are chemical structures in the sense of non-polymeric organic chemistry. Such **structure languages** are

the scientific language of chemists (**chemists' language**), in as far as it is used for describing non-polymeric chemical structures

the formally defined language GENSAL[7] (used as **surface GENSAL** for encoding and displaying structures within the complete GENSAL system for storage and retrieval of generic chemical structures)

the **Extended Connection Table Representation**[8,9] (ECTR, used as an exact internal representation within the GENSAL system)

reduced graphs[13] and **fragment screens**[14] (used as internal representations at a reduced level of information on structural detail in rapid preliminary search operations)

any other system of signs used for the representation of chemical structures (e.g., the surface languages used

in the Markush DARC or MARPAT systems, or the Wiswesser line notation)

Three different entities need to be considered in regard to the **syntax**, **semantics**, and **pragmatics** of structure languages: structural representations (i.e., structural expressions in a wider sense), chemical structures, and users who utilize structural expressions to represent chemical structures.

The **pragmatic** aspect refers to all interrelations between the three entities, i.e., to the interrelations between the user, the structural expression used in a particular operation for some particular purpose, and the structure represented by this expression in this situation. In the GENSAL system, the user may be an encoder using GENSAL as an input language or a searcher using the system for the retrieval of file structures. The use of the GENSAL system is controlled by coding conventions, by conventions on the use of query formulations, and by those basic concepts of representing and retrieving structures that are implemented in the system for the benefit of users. Conventions, however, may prove to be insufficient for practice or they may be applied incorrectly; and the capabilities of internal representations and of search algorithms may also be inadequate for the user's needs. Thus, not only the proper use of GENSAL and of internal representations but also possible deficient use need to be considered as the subject of pragmatics. Furthermore, since a structure to be encoded into GENSAL by an encoder is usually given as an expression in the chemists' language used by a patent applicant, the problem of proper translation from one structure language into another is also a matter of pragmatics.

The **semantic** aspect disregards the user; it refers only to the relation between the structural expression and that structure which is designated by this expression. It is thus restricted to the relation between a structural expression and that structural meaning which is inherent in this expression, regardless of any particular pragmatic situation in which this expression could be used by a user as an exact, or possibly inexact, or even incorrect representation of the structure which is to be represented in this situation.

The **syntactic** aspect disregards the user and the meaning of a structural expression; it refers merely to the grammatical form of an expression, in particular to the grammatical interrelation between partial expressions within a complex expression. The syntax, thus, is restricted to the aspect of grammatically well-formed structural expressions, regardless of any particular pragmatic situation and of any particular meaning associated with these expressions.

In linguistics, the usefulness of the distinction between the semantic and the pragmatic aspect is undisputed. However, the exact delineation of the demarcation line between semantics and pragmatics is controversial in various linguistic approaches. The problem is readily highlighted by Wittgenstein's sentence: "The meaning of a word is its use in language" ("Die Bedeutung eines Wortes ist sein Gebrauch in der Sprache").[15] Nevertheless, even without an exact definition of the demarcation, the explicit distinction of a comprehensive, genuinely pragmatic aspect from the restricted semantic aspect is useful for the analysis, description, and treatment of problems in the use of languages for storage and retrieval of generic chemical structures. In particular, the proper translation of ambiguous structure descriptions, which frequently occur in patents, into the GENSAL language is a genuine problem of pragmatics; it cannot be considered in terms of the syntax and the semantics of chemists' language and of GENSAL alone. In general, the primacy of pragmatic demands is decisive for the design of the GENSAL system; the syntactic and semantic features of GENSAL and of internal representations, search algorithms, and facilities for formulating varying query intentions must be adjusted to the users' needs.

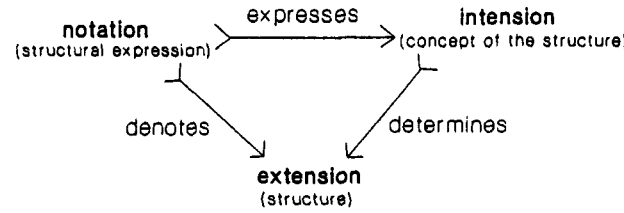

Figure 1. Triangle of meaning-relations.

3. THE MEANING OF STRUCTURAL EXPRESSIONS

In linguistics, there is no commonly accepted model for the theory of meaning that can also be applied cogently to structure languages. In order to discuss the adequate treatment of structural differentiations in the GENSAL system from the pragmatic point of view, it seems useful to differentiate between extensional and intensional facets of meaning. The concept of the intension (as distinct from the extension) has been introduced by Carnap[16] as a development and formalization of Frege's concept of *Sinn* (sense) as distinguished from *Bedeutung* (denotation).[17] The meaning of structural expressions may then be conceived in terms of a "triangle of meaning-relations", which is a special variant of similar triadic models discussed in linguistics in response to Peirce's concept of "semiosis".[18] This semiosis, the constitutive characteristic of a sign as a meaningful sign, is the interaction of three factors: the sign, the interpretation of the sign, and the designated object.

3.1. THE TRIANGLE OF MEANING-RELATIONS

The triangle of meaning-relations (Figure 1) is applicable to expressions in any language. It may readily be explained by examples of simple or complex structural expressions as they occur as formulas or within formulas in patents in organic chemistry.

The **notation** of a **structural expression** (e.g., "tertiary butyl") is the mere syntactic form of the expression (e.g., the sequence of the two words "tertiary" and "butyl"). This notation **denotes** a special **extension**, i.e., it denotes a particular **structure**, and it **expresses** a particular **intension**, i.e., it expresses the **concept** of this structure. Since notations may be fixed by arbitrary conventions, the reference from the notation to the denoted extension does not immediately originate in the notation itself, rather, it is constituted indirectly by the expressed intension which **determines** the extension by being the proper conceptual interpretation of the notation.

In chemists' language, the distinction between a chemical *compound* and its *structure*, which is only a *property* of the compound, is usually not reflected in an explicit manner. (This peculiarity of chemists' language, and the relationship between compounds and structures, is discussed in detail in Section 4.1). In the context of this paper, in any case, the extension of a *structural* expression is always conceived as a *structure*.

3.2. SOME TYPOLOGICAL DIFFERENTIATIONS OF STRUCTURES AND STRUCTURAL EXPRESSIONS

For reasons of conformity with practice in referring to structures in chemistry, it is useful to understand the terms "structure", "structural expression", and "notation" not only as referring to the complete structures of specific compounds but also to partial structures and to classes of complete structures or partial structures. Thus the notations "isobutane",

"tertiary butyl", "branched alkane", and "tertiary alkyl" are structural expressions; they denote extensions which are structures. For structural expressions in organic chemistry, then, the denoted extension is always:

a **specific structure** (i.e., the structure of a specific compound or a part of such a structure) or a **generic structure** (i.e., a class of two or more specific structures)

and it is, at the same time

a **partial structure** (i.e., a "radical", "group", "moiety", "residue", etc.) or a **full structure** (i.e., the complete structure of a single specific compound or a class of such complete structures)

This characterization of a structure—as being a **specific full**, a **specific partial**, a **generic full**, or a **generic partial** structure—can be applied to the extension of any arbitrarily complex expression in organic chemistry. Thus, a complex Markush formula in a patent is a structural expression denoting a generic full structure. This characterization of *structures* may also be transferred to the *expressions* of structures. A Markush formula, then, may be referred to as a generic full expression or a generic full notation.

This characterization of organic structures cannot, of course, be applied simply in the same manner to the extensions of formula expressions in inorganic chemistry, or to the extensions of expressions describing organic polymers, because these expressions, in most cases, are not really *structural* expressions, e.g., "$Na_2SO_4 \cdot 10H_2O$" or "copolymer of styrene and butadiene".

Within a structure language, e.g., within chemists' language in patents, different kinds of notations may be distinguished in regard to their particular syntactic form and semantic functions:

common names referring to compounds ("geraniol", "geranyl")
specific nomenclatural terms ("ethane", "ethyl", "acetyl", "phenyl", "chloro")
generic nomenclatural terms ("alkane", "alkyl", "acyl", "aryl", "halogen")
shortcuts for these ("Et", "Ph")
line formulas ("C_2H_5", "C_6H_5")
structure diagrams, etc.

Several different *kinds* of notation may be combined with one another in one notation for a Markush formula. This generic full expression is then a **composite notation** (or **composite expression**), which is composed of **simple notations** (or **simple expressions**) and special symbols or phrases linking the simple notations syntactically and semantically with one another. By way of example: "C_6H_5-R, R = phenyl or cyclohexyl, both optionally substituted by OH, halogen, or alkyl".

An exact definition of composite and simple notations can be given only for formalized languages. For structure languages in general, a provisional, heuristic definition may be given by reference to the obvious texture of Markush formulas. These composite notations, it is clear, contain syntactically and semantically distinct partial expressions denoting distinct partial structures. Simple notations are partial expressions like "alkyl", "ethyl", and "tertiary butyl", which cannot be subdivided further into parts similarly so that each part still denotes a smaller partial structure. The part, "tertiary", of "tertiary butyl", for example, is not a structural expression because it does not denote a structure but expresses only a particular structural feature within a structural expression. Distinct structure diagrams of full or partial structures should also be referred to as simple notations, in correspondence with specific nomenclatural terms.

Generic structures, i.e., classes of two or more specific structures, may either be finite classes (denoted, e.g., by "C1–4 alkane", "C4–8 tertiary alkyl", etc.) or potentially infinite classes (denoted, e.g., by "alkane", "tertiary alkyl", etc.). Generic full structures and generic partial structures may be denoted not only by those simple or composite expressions which consist of (or which contain) generic nomenclatural terms denoting homologous series ("cycloalkane", "cycloalkyl", "alkylbenzene", "phenyl substituted in the para position by C1–5 alkyl") but also by notations using other means of expressing the alternation of distinct specific structures within a class [e.g., "phenyl, substituted in the para position by methoxy, amino, hydroxy, methyl", "monochlorophenyl", "*p*-phenylene-$(CH_2)_n$-Cl; n = 1–3"]. Thus, "R = methoxy, amino, hydroxy, methyl" is conceived as a generic partial structure in the same manner as "R = C1–5 alkyl".

3.3. THE RELEVANCE OF INTENSIONAL DIFFERENTIATIONS

For structure languages, the intension expressed by a notation is the concept of a structure. From the strictly semantic point of view, the intension may be conceived as the semantic interpretation of the notation as an expression of a particular structure. This semantic interpretation is independent of any mental process of interpretation in any given pragmatic situation; it is *the* commonly valid interpretation of the notation in the respective language. From the more comprehensive pragmatic point of view, a notation is used as a means of communication in a particular situation, and the intension may be identified with the particular information given by this notation in this situation (the term "information" is understood here in the simple pragmatic sense of "a message in a process of communication", rather than in the formal syntactic sense of Shannon's theory of information).

According to the triangle of meaning-relations, the structural expression becomes a meaningful expression by virtue of the relations between its notation, its intension, and its extension. One may say, therefore, that a structural expression "has" a notation, an intension, and an extension. This phrase, however, does not indicate a distinction between a structural expression and its notation; it indicates, as a paraphrase, that a structural expression *is* always a *meaningful* notation, expressing an intension and denoting an extension. With regard to the **meaning** of a structural expression, the intension and the extension of this expression can be said, in paraphrase, to be the intensional and extensional facet of its meaning, or even its intensional and its extensional meaning.

The three binary **meaning-relations** are

the **relation of denoting** between notation and denoted extension
the **relation of expressing** between notation and expressed intension
the **relation of determining** between intension and determined extension

These are many-to-one correspondences, symbolized in Figure 1 by the directed arrows: one extension may be determined by two or more intensions, and each of these intensions may be expressed by two or more notations, so that one extension may be denoted by two or more notations.

In chemists' language, for example, the three different notations "γ-butyrolactone", "α-pyrrolidone", and the corresponding structure diagram express three different intensions but denote the same extension: the same structure is conceived and determined differently in different scientific contexts as a cyclic amide, as a heterocyclic ketone, or as the bare aggregation of atoms connected by bonds. From the pragmatic viewpoint of storage and retrieval of structures, the extension of a structural expression is usually of primary interest, the notation and the intension being merely incidental means of denoting and determining the extension. The translation of

these three different notations of chemists' language into GENSAL, therefore, is one and the same structure diagram, and the translation into the ECTR is one and the same connection table. For pragmatic reasons, the disregard of different intensions determining identical extensions is usual and useful for notations denoting *specific* full or *specific* partial structures for purposes of storage and retrieval of structures.

In some special cases of *generic* full and *generic* partial structures, however, even intensional differences for identical extensions need to be taken into account even for computer storage and retrieval of structures, again for pragmatic reasons. The class of halogen atoms, for instance, may be denoted in Markush formulas by two different structural expressions, by "R = halogen" and by "R = F, Cl, Br, I" (neglecting the halogen element astatine, which is not included as a halogen in organic chemical practice). The generic partial structure denoted is a finite class of specific partial structures, which are single atoms in this particular case. The two different notations denote the same extension by expressing two different intensions which determine this extension in two different ways: generically as a whole class of specific substituents having the common property of being halogen atoms and specifically as a set of four different specific substituents which are enumerated as specific partial structures.

This differentiation may seem to be purely academic. For these and similar examples of different notations denoting the same generic extension, one might question first whether really different intensions determining the same extension (i.e., different concepts of the same generic structure) are expressed at all, and moreover whether the distinction between these different intensions is relevant to the practice of structure retrieval. These questions can only be decided pragmatically on the basis of criteria set by the users' needs with regard to the capabilities of the retrieval system used.

Failure to observe the purely intensional difference between the two extensionally identical expressions "R = halogen" and "R = F, Cl, Br, I" may, in practice, result in undesired consequences. In the retrieval system, both expressions could be encoded and stored in the same specifically describing manner, i.e., by enumeration of the four specific halogen atoms. However, seachers requiring exclusively the specifically described partial structure iodine, which is comparatively infrequent, in response to a particular part of the query (e.g., for substitution on phenyl), would also retrieve all those rather frequent yet unwanted generic partial structures which are given originally in a patent by "halogen" (e.g., in "phenyl, optionally substituted by halogen"). Vice versa, it would not be adequate to encode and store "R = F, Cl, Br, I" (or even "R = F, Cl, Br, I, At") in the same generic manner as "R = halogen", viz., by a suitable generic representation for all halogen atoms. In a search for the specifically described iodine (or even for astatine), the user again could not distinguish between the different intensions, although he regards the explicit enumeration as the more relevant expression.

For pragmatic reasons, therefore, the distinction between the intension determining a finite class generically by a generic term and the intension determining a finite class specifically by enumeration of the specific structures must be made in input operations. In a search, too, the system should enable the user to make this distinction if he considers it a necessary distinction.

For similar pragmatic reasons, the same distinction should be made for a finite class that is given as a subset of a potentially infinite homologous series. The generic partial structure, which is the class of all specific alkyl radicals with one to four carbon atoms, for instance, may be determined by two different notations: "R = C1–4 alkyl" instead of "R = methyl, ethyl, *n*-propyl, isopropyl, *n*-butyl, isobutyl, *sec*-butyl, *t*-butyl". The first intension conceives and determines the denoted class as a homologous series with a restricted range of carbon atoms; the second intension conceives and determines the same extension in a different manner as a set of eight distinct specific partial structures. (In accordance with the triangle of meaning-relations, each of these two different intensions may also be expressed by several other notations, the first, e.g., by "alkyl with fewer than 5 carbon atoms", or the second by the enumeration of distinct structure diagrams instead of distinct nomenclatural terms.) Here again the disregard of the intensional difference would make impossible, in searches for specifically described partial structures, e.g., for *t*-butyl, the necessary differentiation from generically described subsets of alkyl.

In all cases where the user might wish to differentiate in searches between different intensions determining the same extension, the difference in intension is seen by the user as a difference of relevance to the particular search purpose he has in mind. Not infrequently, the user simply wants to divide a search to get two answer sets: the more relevant answer set first, and after that, and only if necessary, the set of less relevant answers. For the two examples discussed, the difference of relevance originates in the different degree of specificity of determining a finite class.

There are also other merely intensional differences that might be seen as differences in relevance for particular search purposes:

> In a search for tertiary butyl and classes including it, the expression "alkyl, optionally tertiary" might be regarded as more relevant than the extensionally identical "alkyl".
>
> In a search for alkyl radicals with more than 25 carbon atoms, "alkyl, optionally with more than 20 carbon atoms" might be regarded as more relevant than the extensionally identical "alkyl".

In these cases, the difference in relevance originates in an emphasis on particular structural features which are expressed explicitly, in a redundant and overemphatic manner, although they are expressed implicitly in any case. Whether this special type of intensional differentiation should be searchable is a matter of pragmatics. It could be made searchable simply by additionally encoding the appropriately specified generic nomenclatural term as alternatives, e.g., "alkyl/*t*-alkyl" or "alkyl/alkyl ⟨20-⟩", with a special proviso: these alternatives must be distinguishable by appropriate search algorithms, i.e., the user must be able to define the level of matching needed in such a way that "alkyl" is excluded from the retrieved answers as too high a generalization, whereas "*t*-alkyl" or "alkyl ⟨20-⟩" is included as an acceptable level of generalization. The possibility of implementing this search feature in the GENSAL system is discussed in the following paper in this issue.

The *extension* of an exact structural expression, e.g., of "*t*-butyl", is context-independent, i.e., the bare structure is denoted exactly and exclusively by the isolated notation. The *intension* of an exact structural expression is also mainly expressed by the notation, but it may be influenced by the context, even in a formalized language like GENSAL. From the pragmatic aspect, the complete intension of a structural expression, i.e., the fully understood concept of a structure, is not restricted to the mere *structural* information on the structural properties of the isolated structure. It also includes *operational* information, i.e., it comprises, more generally, the entire information about the structural meaning and the appropriate use of the expression within its particular semantic and pragmatic context. The intension, being expressed explicitly by the notation and implicitly by the context, determines how the expression is to be understood as an expression

of a structure and how the expression is to be used in operations of communication or computation. The intension, thus, cannot be seen from a restricted aspect focusing exclusively on the semantic function of the isolated notation; it must be seen from the more comprehensive pragmatic aspect, comprehending the pragmatic function of the expression within its particular pragmatic context.

The operational aspect of intensions becomes clearly visible in a comparison of two identical notations, e.g., "*t*-butyl", when embedded in the different contexts of a file record or a query formulation. The two contexts differ from one another

- syntactically (in GENSAL, e.g., by "INPUT" and "QUERY", being associated with the full file and query expression)
- semantically in the nonstructural part of the intension: the difference in meaning is just the difference between an intension determining a structure to be operated as a file structure and another intension determining the same structure to be operated as a query structure
- pragmatically in its use, which is clearly different for the same notation in file and query

The different influence from different intensional contexts may cause intensional differences which might be seen, again, as differences in relevance with regard to particular search purposes. In the following cases, extensionally and even notationally identical expressions that are embedded in different intensional contexts might be seen by the searcher as having differing relevance:

- "R = hydrocarbyl EG *t*-butyl", indicating in GENSAL that the generic nomenclatural term "hydrocarbyl" is instantiated by "*t*-butyl" in an example of the patent
- "R = alkyl PREF *t*-butyl", indicating in GENSAL that *t*-butyl is the preferred instantiation of "alkyl"
- the structure diagram of "*t*-butyl", occurring as a partial expression within the constant part of a Markush formula

Such context-dependent intensional differences of extensionally identical expressions usually become visible only in displayed answer sets of file structures. However, the cases discussed above can be distinguished, if this is really desired, because of the obvious syntactic differences ("EG", "PREF", structure diagram in constant part). This possibility will be discussed in Section 5.3.

4. COMPARATIVE ASPECTS OF STRUCTURE LANGUAGES INVOLVED IN THE USE OF THE GENSAL SYSTEM

GENSAL is a formalization of those parts of the language of chemistry that are used for describing organic structures. Problems in the use of GENSAL originate just in this derivation from and correspondence with chemists' language. Since GENSAL is intended to be used as a computer-processable and yet easily applicable substitute for chemists' language, its expressions must combine exactness with the best possible adaptation to the syntactically versatile and semantically sometimes rather inexact expressions in patents. These requirements result in special features of syntax, semantics, and pragmatics of GENSAL as compared with chemists' language and other structure languages.

4.1. PECULIARITIES OF CHEMISTS' LANGUAGE

Some peculiarities in the use of chemists' language, which are also relevant to the use of GENSAL and to a comparative discussion of structure languages, are briefly sketched here.

In everyday communication in organic chemistry, a chemical *compound*, its *structure*, and the *formula* of this structure usually need not be distinguished from one another in a terminologically explicit manner. Without being misunderstood, a chemist may point to a structure diagram and say: "This is an analgesic". This usage is the consequence of a phenomenon which seems to be singular in science: it is only in organic chemistry that the most important subjects of this science (viz., the chemical compounds) are uniquely identifiable and, in many aspects of their behavior, sufficiently characterizable by *one single property* (viz., by the structure); it is only in chemistry that such a uniquely identifying property can be described in a very simple language (viz., by a structure diagram) consisting basically of only two kinds of simple symbols (viz., symbols of atoms and of bonds), by combining the descriptions of partial properties (viz., of partial structures). Consequently, the systematic nomenclature of chemical *compounds* mainly refers to the *structure* of compounds, and most nomenclatural names may be used and conceived as names of *compounds and* as descriptors of *structures*.

This singular situation leads to special usages in chemists' language. These usages, blurring the distinction between compounds, structures, and structural expressions, are sufficiently exact in the everyday language of scientific publications and patents, where chemists' language is used as an **object language**, i.e., as a means of talking about nonlinguistic objects, namely **compounds** or **structures**. But in considering the syntax, semantics, and pragmatics of structure languages, a theoretical **metalanguage** is used as a means of talking about linguistic objects, i.e., about the form, the meaning, and the use of *expressions* of a language. Thus, the object language itself, its objects, and the pragmatic circumstances of its use, are made the objects of a more comprehensive metalanguage. For discussions in such a metalanguage, a disregard of the difference between compounds, structures, and structural expressions would be confusing. In particular:

> In chemists' language the name of a *compound*, (e.g., "geraniol") is conceived—if used in a scientific context focusing on genuinely structural topics—as a name of a *structure* (i.e., as a structural expression). And conversely, the description of a structure (e.g., the structure diagram of "geraniol")—when used in a context focusing on nonstructural properties of compounds—is often used as a name of a compound. But for discussions of structure languages in a metalanguage, the difference between compounds and structures cannot be ignored. In considering the peculiarities of chemists' language and the problem of translating expressions of chemists' language into GENSAL, for instance, this difference is reflected as a difference in the treatment of structure diagrams, denoting structures by immediate description, and expressions like "geraniol" or "geranyl", denoting structures indirectly by reference to the known structure of a compound.
>
> In chemists' language, structure diagrams of specific structures are conceived as graphic models of these structures. Operations on these structures, e.g., comparison of structures for identity or substructural embedment, are performed as visual operations on these models, i.e., on expressions of structures. Similarly, such operations are also performed on verbal expressions, e.g., on "phenyl" and "benzyl", by explicit or implicit reference to the corresponding structure diagrams. Thus, *expressions* of structures are treated and conceived as if they were *structures*. However, the distinction between structures and structural expressions (i.e., between extensions and notations) is essential for discussion of structure languages in a metalanguage, e.g., for considering the fact that one structure may be denoted by several different notations, or for discussing the circumstance that one notation in GENSAL may be the

proper translation of several notationally different yet extensionally identical expressions of chemists' language.

Similarly, a distinction between a class of compounds, the corresponding generic full structure, and the corresponding generic full expression must be made in a metalanguage, in analogy with the distinction between a compound, its specific full structure, and the corresponding structural expression.

Furthermore, the distinction between full and partial structures and the distinction between specific and generic structures, which are sometimes blurred in chemists' language, should always be kept in mind:

> Since operations on specific *full* structures are very similar to operations on specific *partial* structures and since specific *full* and specific *partial* structures may be denoted in a very similar manner ("ethane–ethyl", "C_2H_6"–"C_2H_5"), they are designated and conceived uniformly as *structures* in chemists' language. This egalitarian treatment should be admitted in the metalanguage too, if the context or an explicit qualification makes clear whether partial or full structures, or both, are meant.
>
> Since *specific* full and *generic* full structures may be denoted in chemists' language in a formally similar manner (viz., by combining expressions for specific partial structures, which in generic formulas must be marked by syntactic means only as being constant or variable: "C_6H_5–OH"—"C_6H_5–R, R = OH, ethoxy, dimethylamino") and since instead of *specific* partial expressions, formally very similar *generic* partial expressions may be contained in generic formulas ("C_6H_5–R, R = OH, alkoxy, dialkylamino"), *classes* of specific full or specific partial structures are also designated and conceived as *structures*. This egalitarian treatment should be admitted in the metalanguage too, if it is made clear whether specific or generic structures, or both, are meant.

As discussed above, the relation between a compound and its structure can be seen as a one-to-one correspondence. This one-to-one correspondence is used as a pragmatic principle, in systematic classifications and in retrieval systems, for identification and retrieval of compounds by means of their structures. In regard to its factual use, this principle must be defined, more precisely, as a one-to-one correspondence between an individual **chemical entity** and its structure. This chemical entity may be a compound in the usual sense or a distinct "form" of a compound, e.g., the D-form of lactic acid.

Deviations from this principle seem to be given in special cases, which could be conceived as cases of a one-to-many correspondence. These are, for example, cases of mesomerism with different "resonating structures"; cases of equilibria between tautomeric or other isomeric structural forms (e.g., ring-chain forms of sugars); or cases of different, differently expressed structural conceptions of organic salts, addition compounds, complexes, etc. Such cases may cause some difficulties in retrieval systems, in particular, if they are really treated as cases of a one-to-many correspondence, i.e., as the occurrence of *one* chemical entity in the form of *different* structures. This treatment, however, is neither cogent by reason of facts nor useful for pragmatic reasons. Therefore, in the practice of retrieval systems, these difficulties are usually diminished or eliminated, in a pragmatic manner, just by using the conception of a one-to-one correspondence as a pragmatic postulate. That means since all physical or chemical diferentiations of molecular species strictly correspond to structural differentiations of molecules, every differentiation of molecular species into distinct, individual chemical entities according to pragmatic needs can be conceived and treated, in principle, in such a manner that one distinct entity corresponds to one structure and may be identified by this structure. ("Structure" in this sense is to be understood as that form of constitution, configuration, or even conformation which corresponds to the respective, pragmatically defined entity.)

Examples of the application of this pragmatic postulate to problematic cases:

> If a case of mesomerism is described by the formulas of two or more fictitious "resonating structures", e.g., the mesomerism of potassium rhodanide by "$[^-S—C{\equiv}N \leftrightarrow S{=}C{=}N^-]K^+$", then this is to be conceived as a special kind of notation denoting *one* mesomeric structure. In retrieval systems, anyway, cases of mesomerism are usually denoted—according to special structuring conventions, which are factually notation conventions—by using only the notation of one of the different resonating structures (e.g., one of the two possible arrangements of alternating double bonds in a structure diagram of ortho chlorophenol), or by using special notational means (e.g., normalized bonds for the mesomerism of benzene rings in the CAS system[19]).
>
> Similarly, different notations might be fixed by different structuring conventions in different retrieval systems for the retrieval of the compound sodium acetate (e.g., the line notations "CH_3COONa", "$CH_3COO^-Na^+$", "$CH_3COOH{\cdot}Na$", the corresponding structure diagrams, and the corresponding internal representations). They are to be conceived merely as *different notations* of one and the same extension, viz., of the *one structure* of the chemical entity sodium acetate—even if this structure is to a certain extent a conceptual abstraction by neglecting the factual difference which exists for the physically structural arrangement of sodium cation and acetate anion in different states of the compound, e.g., in the solid state and aqueous solution. (Of course, if in a scientific context, or even in a special retrieval system, for pragmatic reasons, different states of sodium acetate are conceived and treated as different entities having different structures, then "CH_3COONa" and "$CH_3COO^-Na^+$" may denote different extensions.)
>
> Tautomers may be treated—according to the special pragmatic needs which are given by the chemical facts or by the technical feasibilities of the retrieval system—as *different* chemical entities with *different* structures (e.g., keto-enol tautomers in the CAS system), or as one chemical entity whose one structure is denoted by the use of special notational means and conventions (e.g., tautomers according to the restricted CAS definition, which are denoted internally by means of normalized bonds[19]).

4.2. TRANSLATION AND TRANSFORMATION

In the linguistics of natural languages, the differentiation between the concepts of translation (from one language into another) and transformation (within one language) is unproblematic, provided that the usual concept of distinct standard languages is accepted. The proper application of these concepts to structure languages, however, seems to be problematic: structure languages are artificial, more or less formalized languages, they are intricately interrelated with one another, and they show in parts similar syntactic and semantic features. On the one hand, therefore, it is not quite clear what should be regarded as a distinct structure language at all. On the other hand, translations and transformations are, for these languages, very similar procedures of converting expressions, and they can seemingly be distinguished only by the criterion of whether these expressions belong to one and the same or to different structure languages. Thus, it seems that for structure languages the concepts of translation, or transfor-

mation, and of a distinct language can be defined only by a circular definition.

As matters stand, these concepts should be applied to structure languages quite pragmatically (viz., with respect to the pragmatic context of the problem to be discussed) and quite naively (viz., in analogy to the applications of these concepts to natural languages). So one may state:

> It is a characteristic of **translations**, in general, that there may be difficulties with the exact translation of the meaning of particular expressions from one language into another. The expression conjugated alkadienyl of chemists' language, for instance, cannot be translated exactly into GENSAL if the structural feature of conjugation of multiple bonds cannot be expressed in this language. Nevertheless, there will be a proper translation, i.e., a best possible translation, which is identical with the translation of the extensionally and intensionally broader expression alkadienyl. Thus, in translations, not only intensions but also extensions cannot always be retained exactly.
>
> For **transformations** of structural expressions within one language, however, the exact extension must always be retained, and the intension should be retained where appropriate. So the structure diagram of trifluoromethyl and the extensionally and intensionally identical notations "trifluoromethyl" and "CF_3" are interchangeable in transformations within GENSAL; "α-pyrrolidone", "γ-butyrolactam", and the corresponding structure diagram, which are intensionally different yet extensionally identical expressions, may be interchanged with one other within chemists' language if this is appropriate in the particular scientific context, but the extensionally identical yet intensionally different expressions "halogen" and "F, Cl, Br, I" must be treated in the GENSAL system as noninterchangeable expressions for the pragmatic reasons discussed in Section 3.3.
>
> It is a characteristic of a **distinct language** that it provides the user with sufficient syntactic and semantic facilities for talking about the objects of this language in communicative and meaningful texts (the term "text" being understood in a wider sense, including also those aggregates of expressions which form descriptions of full structures in structure languages). Conversely, it is also a characteristic of a distinct language that every coherent meaningful text, talking about any object(s), contains exclusively expressions belonging to one and the same language. [The fact that a text discussing structure languages contains the expressions of *different* structure languages, by the way, is no objection: it is a characteristic of the comprehensive *meta*language that it contains the object language(s) discussed as a part.]
>
> For the encoder using GENSAL as an input language, consequently, chemists' language is not a mixture of different languages. It is a *single* language, because all those different forms of notations that are comprehensible to chemists may be contained in one text, i.e., in one Markush formula. (In another pragmatic context, e.g., if the problem is the transcription of exclusively nomenclatural names into structure diagrams, or vice versa, then this may be conceived as a translation from one distinct language into the other—in particular if the typical problems of proper translation occur.)
>
> Analogously, the different forms of notations in GENSAL are not different languages because they may be contained within one GENSAL expression of a generic full structure.
>
> Two languages may be very similar to one another by having large parts of syntax and semantics in common. But they are, nevertheless, different distinct languages if there are any particular syntactic or semantic features that are admitted exclusively in one of these languages but not in the other. GENSAL, therefore, would remain a language different from the syntactically and semantically versatile language of chemists—even if GENSAL became a commonly accepted part of chemists' language.

4.3. SYNTAX AND SEMANTICS OF GENSAL AND OF THE ECTR

The syntax of GENSAL, as the syntax of a formalized language, is well-defined: the generation of correct expressions is controlled by strict rules, which are defined in a suitable manner, e.g., by syntax diagrams.[7] Within the GENSAL system, the expression of a generic full structure in GENSAL is input by means of a suitable input system. This expression is translated by the GENSAL interpreter into the corresponding internal representation, namely, the ECTR. (Modifications which have been made to the original specification of the ECTR are discussed in the third (of three) paper in this issue.) By virtue of this translation, the ECTR is derived from the GENSAL expression but is generated according to its own well-defined syntax. The ECTR and GENSAL, therefore, are to be conceived as different distinct languages.

In translating Markush formulas into generic full expressions in GENSAL, the notational versatility of chemists' language cannot be retained—if only because of economic and technical reasons. In GENSAL, therefore, only a restricted set of standard text terms consisting of particular generic nomenclatural terms, specific nomenclatural terms, shortcuts, and line formulas is allowed. Partial expressions drawn as structure diagrams in GENSAL are translated by the interpreter directly into partial connection tables in the ECTR. Specific text terms recognized in the dictionary are also translated into partial connection tables in the ECTR, but they are translated indirectly by means of the dictionary containing the specific text terms together with the corresponding partial connection tables. Generic nomenclatural terms like "alkyl" are also translated by the dictionary, which contains the generic nomenclatural terms and the corresponding, appropriately defined parameter lists.[14]

Expressions like "electron-withdrawing group", having no definite structural meaning, are an exception. Since they cannot be translated into connection tables or parameter lists, they are incorporated into GENSAL and the ECTR as text strings, being referred to as "Other Terms". They are alien elements in syntax and semantics of both languages and are discussed in more detail in Section 4.5.4.

The restriction of the number of allowed standard text terms is not a restriction of the expressive power of GENSAL per se, because infrequent specific expressions like "geranyl", which are not contained in the dictionary, are translated by the encoder into the extensionally identical structure diagrams. Infrequent generic text terms like "bicyclic cycloalkyl" are translated by the encoder into an allowed generic nomenclatural term, which is supplemented by the appropriately restricting parameter values. Thus, the expressive power of expressions denoting generic radicals in GENSAL is not restricted by the number of allowed generic nomenclatural terms, but rather by a restricted set of parameters, which can represent only a restricted number of structural features of generic radicals.

The structural meaning of any particular expression in GENSAL (Other Terms excepted) is fixed by its translation into a connection table or into a parameter list. Within the limits of this formalism, thus, not only the syntax but also the

semantics of GENSAL and of the ECTR are well-defined. There may be restrictions, however, with regard to the expressive power of these structure languages. Limitations may be given, for instance, by the inability of the particular type of connection table used to represent particular features of specific structures exactly (e.g., isotopes, stereochemistry, charges, tautomerism, aromaticity, etc.), by the inevitably restricted expressive power of a finite set of parameters, or by possibly inadequate parameter values chosen for defining the parameter list of an allowed generic nomenclatural term. In principle, however, all these restrictions can be reduced or eliminated by appropriate amendments.

Some particular features of the syntax and semantics of the ECTR are of particular relevance to search algorithms. In organic structural chemistry, the special reation between a generic structure (i.e., a class of specific structures) and a specific structure that is a member of this class is usually recognized intellectually by comparing these structures by means of their expressions. For ECTRs, this comparison of query and file structures can be performed by algorithmic operations on partial connection tables and parameter lists (which are, more precisely, lists of parameter *values*). The algorithm might, for instance, be based on the recognitive capability of a "chemical grammar", e.g., TOPOGRAM,[20] but here we consider the possibility of a "specific" parameter list, derived from a connection table, which can then be compared simply with the "generic" parameter list representing a generic radical. Furthermore, different levels of matching, e.g., the difference between matching "ethyl" with "C1–4 alkyl" and matching "ethyl" with "hydrocarbyl", can optionally be distinguished, according to user-defined gradations, by simple formalisms in search algorithms as described in the following paper.

A precondition for the exactness of these formalisms is that the exact definition of all structural features which are used as standard parameters must be exact in an operational sense: it must be definitely decidable in an algorithmic procedure, for any connection table, whether this structural feature is absent (parameter value ⟨0⟩) or how many times it is present (parameter values ⟨1⟩ or ⟨2⟩ or ⟨3⟩ etc.). On the basis of such an algorithm for each standard parameter, **specific parameter lists**, containing exclusively *distinct integers* as parameter values, can be derived from connection tables of specific structures. Equally, **generic parameter lists** of generic nomenclatural terms, containing *ranges of integers* as parameter values, can be defined by considering, for each parameter, the range of values that is given by all of the specific structures of a homologous series. In the GENSAL system, thus, the internal representations originate from the fundamental concept of the molecular graph, but the comparison of generic nomenclatural terms with one another or with specific expressions is performed by using the derivative concept of parameter lists. Both concepts are used in an exact manner.

The exactness of the well-defined semantics of GENSAL and of the ECTR is not exactness in the sense of an exactly differentiating representation of even the finest structural details, it is the definiteness of an exactly regulated syntax. Each of the languages, GENSAL and the ECTR, is a syntactically and semantically restricted language, but within its limits it is (or, at least, it is intended to be) a definite language with respect to its definitely regulated syntactic and semantic formalisms. That means, in particular:

> Within the expression of a generic full structure in GENSAL or within an ECTR, the partial expressions are linked to one another in an exact way and according to definite rules.

Within expressions of GENSAL, correct transformations can be performed according to definite rules on mere notations, irrespective of any intellectual grasp of the meaning. A specific nomenclatural term, for example, can be transformed via the internal dictionary into an extensionally identical structure diagram, or vice versa, or it can be transformed into another specific text term with an identical extension. In practice, such transformations are not performed during computer operations, but in principle, they could be performed algorithmically, and they indicate the definiteness of GENSAL with respect to syntactic and semantic operations. In practice, such transformations are used implicitly by encoders as a flexible means of choosing the most convenient expression allowed in GENSAL, e.g., "tosyl" instead of the corresponding structure diagram, or "CF_3" instead of "trifluoromethyl".

Translations from GENSAL into ECTRs are performed by the interpreter according to definite rules on notations. Thus, the translation is syntactically definite, i.e., an exact conversion from one syntax into another, although it may not be completely definite as a semantic procedure (i.e., as a transfer of meaning) for certain kinds of expressions. (Conditions which narrow the extension, e.g., "IF...THEN...", can be formulated in GENSAL, but they are neglected during translation by the current version of the interpreter; the translationally corresponding ECTR, then, is a semantically inexact translation because of its broader extension.) This occasional slight inexactness of translations from GENSAL into the ECTR does not result in incomplete recall, but it may be seen as a reduction of semantic precision, causing false drops.

Within GENSAL and the ECTR, it is possible to decide definitely whether or not two expressions (or, correspondingly, the structures denoted by these expressions) match, by syntactically and semantically definite operations, and which particular matching-relation exists between these structures. The **matching-relations** (e.g., identity, inclusion, intersection, as discussed in the following paper) are to be conceived as genuine class relations between classes of specific structures, provided that a single specific structure is also conceived as a class, viz., as a class with only one member. A match is determined, then, if two classes contain identical members; if one class includes the other; or if two classes have at least one member after intersection. This matching comparison can be performed for partial and for full expressions (structures). A match between two complex generic full expressions (structures) may be seen as a path (**matching-path**) of matching partial expressions (structures) through the two generic full expressions (structures).

Similarly, it is possible, within both languages, to decide whether or not the special matching-relation of substructural embedment exists, by means of syntactically and semantically definite operations. This special relation is not a relation between classes but between molecular graphs. It may be combined with class relations.

These matching comparisons can be performed algorithmically in ECTRs according to relatively simple formalisms. In principle, however, they could also be performed algorithmically in GENSAL, viz., by implicit reference to the translationally corresponding ECTR, without any explicitly intellectual comparison of meanings.

Thus the syntactic definiteness of translations from GENSAL into the ECTR and the definiteness of syntactic and semantic operations within each language make algorithmic transformation, translation, and comparison (with regard to matching) of expressions possible. This operational definiteness, which is the basis for computerized input and search operations, is a **definite decidability** from the pragmatic point of view:

> For any arbitrary expression, it can be decided definitely whether it is a syntactically correct expression of GENSAL or of the language of the ECTR or not.
>
> For any arbitrary ordered pair of correct expressions within one of these languages, the question of matching-relations can be decided definitely. These decisions are extensionally definite, i.e., they are always definite with regard to extensions. Analogously definite decisions on the nonmatching of intensions (while the respective two extensions match) can, in principle, be made only when the difference of intensions is clearly expressed by syntactic means (e.g., by the syntactic difference between "halogen" and "F/Cl/Br/I"), which are used for differentiating between intensions according to pragmatically fixed rules.
>
> For any arbitrary transformation of expressions in GENSAL or in the ECTR, it can be decided definitely whether it is a correct transformation with regard to syntax and extensional semantics. Whether an extensionally correct transformation is also intensionally correct can be decided under similar conditions as discussed for nonmatching intensions.

4.4. SYNTAX, SEMANTICS, AND PRAGMATICS OF CHEMISTS' LANGUAGE

In comparison with GENSAL or the ECTR, the syntax of chemists' language is not strictly regulated. In a complex Markush formula in a patent, many different forms of notations may occur, including natural language phrases. In cases of such structural expressions, it is impossible to perform correct transformations by definite rules on notations alone, without any intellectual form of assessment of meanings. Seen as a whole, therefore, the syntax of chemists' language is not a system of definite rules which can be used in computer operations. It shows syntactically definite features only in parts, within a particular standardized form of notation or even between two of these forms (e.g., between strict nomenclatural terms and structure diagrams).

Similarly, the semantics of chemists' language as it is used in practice, seen as a whole, cannot be regarded as a system of completely definite relations between unambiguous expressions and exactly denoted structures. Not infrequently, in patents and even in scientific publications, the exact structural meaning of structural expressions remains indefinite even after evaluation of the context.

In principle, however, there should be no problems with indefinite meanings of structural expressions in chemists' language, because this language sets the standards of exactness in describing structures:

> All structural differentiations that can be perceived (through investigation) or conceived (by imagination) exactly in a chemist's *mind* can also be expressed exactly in chemists' *language*, because this language can always be adapted to new differentiations if need be.
>
> The bounds of semantic definiteness are not set by particular limitations of conventional chemists' language, but by the general limitations of human understanding and of human capacity of rational communication by semiotic systems (i.e., by symbols within standardized notations, by diagrams, by words, etc.).
>
> By virtue of the great syntactic and semantic flexibility and by the great variety of structural expressions in conventional chemists' language, a chemist can describe a structure in as definite a form as he wishes it to be, in almost every situation.
>
> There may be particular "new" differentiations of structural details or of structural concepts which have not been known or conceived before and which, therefore, did not "exist" for chemists before. They cannot be expressed in conventional chemists' language. But even then a chemist may arbitrarily extend chemists' language to provide means of describing these new differentiations (e.g., by new terms, new symbols in structure diagrams, or new alphanumeric notations).

The capabilities of chemists' language are not always utilized fully in practice. In patents, for pragmatic reasons, the applicant is sometimes not interested in perfect semantic exactness. In scientific publications, semantic indefiniteness does not usually originate in sloppiness or deliberate inexactness, but it may also originate in a particular pragmatic situation.

> From his pragmatic point of view, the *author* of a publication (or a patent) wishes only to describe exactly those differentiations of structures and structural features that are relevant in the scientific context of his publication (or in the legal context of his patent application). For him, the degree of semantic exactness attained in the structural expressions of his text is pragmatically sufficient.
>
> Also from the pragmatic point of view of the normal *reader*, i.e., of a research chemist (or of a patent agent), these expressions are sufficiently exact in most cases.
>
> In many cases, however, they are not sufficiently exact from the pragmatic viewpoint of an *encoder* who has to translate these expressions into GENSAL. In this language, the capability of expressing even the finest structural differentiations is limited, but those differentiations that can be expressed must be expressed exactly. The encoder, thus, is in the paradoxical pragmatic situation that he cannot express in GENSAL exactly that special semantic inexactness which is given in some expressions of chemists' language.

4.5. DEFINITENESS AND INDEFINITENESS OF STRUCTURAL EXPRESSIONS

Structural expressions and operations on these expressions may be definite or indefinite in several different respects. Different kinds of expressional definiteness and indefiniteness, therefore, can be distinguished as different properties of structural expressions.

4.5.1. MEANING-DEFINITENESS IN THE GENSAL SYSTEM

In structure languages the relations between notations, intensions, and extensions of structural expressions are many-to-one correspondences according to the triangle of meaning-relations. In GENSAL, if we restrict the discussion to those genuine GENSAL expressions which are not Other Terms, the meaning-relations are used as syntactically and semantically definite formalisms. These formalisms can be conceived in a precise manner as functions in the sense of mathematical logic, according to the definition given by Church.[21] The binary relation of denoting, for example, is a function—more explicitly, a *one*-valued *singulary* (or *unary*) function—from arguments that are notations of GENSAL to values that are extensions. Application of this function to any single notation (argument) yields *one* extension (value) only. In accordance with its character of constituting a many-to-one correspondence, this function, on application to several dif-

ferent notations, may yield one and the same extension. The function itself may be conceived as the entirety of syntactic and extensional semantic rules of GENSAL. The three functions forming the triangle of meaning-relations, thus, form a system of definite syntactic rules and of definite semantic rules on notations denoting extensions by determining them through expressed intensions. Since these one-valued functions are aimed at one extension as value, they constitute GENSAL as an extensionally definite system of structural expressions.

From the syntactic and semantic points of view alone, this system is completely independent of any use in practice. Every notation in GENSAL, if regarded as being an **abstract notation (abstract expression)**, i.e., as being isolated from any syntactic, semantic or pragmatic context, denotes its **inherent extension** and expresses its **inherent intension** according to the system of syntactic and semantic rules which constitute GENSAL as a formalized structure language. And every abstract notation, regardless of whether it is a full or partial, a specific or generic, or a simple or composite expression, denotes its inherent extension and expresses its inherent extension exactly, with the sole exception of Other Terms and composite expressions containing Other Terms. This definiteness of the inherent meaning is a property of the notation with respect to GENSAL; it may be referred to as **(inherent) meaning-definiteness**. Structure diagrams (e.g., of "*t*-butyl") and standard text terms (e.g., "*t*-butyl", "C_2H_5", "methoxy", "MeO", "alkadienyl", or "aryl") are **(inherently) meaning-definite** expressions *with respect to GENSAL*. [As will be discussed later (Section 4.5.5), the notation "aryl" is inherently meaning-*in*definite *with respect to chemists' language*.]

Analogously, connection tables and parameter lists in the ECTR are meaning-definite expressions with respect to the ECTR.

As discussed in Section 3.3, the **factual intension** of a **concrete notation (concrete expression)** in the syntactic, semantic, and pragmatic context of a concrete pragmatic situation may be influenced by the context to a certain degree, which sometimes cannot be understood and treated as an *exact* shift of the intension. The extension of concrete expressions in GENSAL, however, is context-independent; the **factual extension** of a concrete notation is understood and treated as identical with the inherent extension in any pragmatic context. The inherent meaning-definiteness of expressions in GENSAL and in the ECTR is therefore a kind of definiteness that is relevant to the use of concrete notations in practice by virtue of this identity of inherent and factual extensions, although it primarily reflects only the syntactic and semantic aspect of abstract notations in GENSAL. (The factual extension may be *different* from the inherent extension for inherently meaning-indefinite expressions *in chemists' language*, as discussed later in Section 4.5.5.)

4.5.2. REPRESENTATION-DEFINITENESS AND TRANSLATION-DEFINITENESS

In practical applications of the GENSAL system, both the factual (=inherent) extension **denoted** by a concrete expression and the extension (i.e., structure) that is to be **represented** as exactly as possible in a given pragmatic situation by an expression in GENSAL and in the ECTR are of interest. From the pragmatic viewpoint, therefore, a second kind of expressional definiteness needs to be considered. It is called **representation-definiteness (representation-indefiniteness)**.

Concrete expressions in GENSAL and the ECTR may be meaning-definite and **representation-indefinite** at the same time, if particular structural features cannot be expressed in these languages. The conjugation of double bonds is a case in point, because this feature cannot be represented in the current version of GENSAL:

> The generic partial structure, which is the class of all "alkadienyl" radicals with two *conjugated* double bonds, then must be represented in GENSAL by "alkadienyl". Seen from the purely syntactic and semantic points of view, this concrete notation "alkadienyl" is invariably—in this as in any other pragmatic situation—a meaning-definite expression with respect to GENSAL; and in this particular pragmatic situation it is at the same time, seen from the pragmatic point of view, a representation-indefinite expression *with respect to GENSAL and to the structure represented* (or, more explicitly, to the structure which is to be represented). In this pragmatic situation, it is an **inexact representation**. But it is, nevertheless, a **proper representation**, i.e., the best possible representation.
>
> Analogously in this pragmatic situation, the parameter list of alkadienyl is representation-indefinite with respect to the ECTR and to the structure represented.
>
> If, in a different pragmatic situation, the extensionally broader class of *all* "alkadienyl radicals" (with two isolated *or* conjugated double bonds) is to be represented in GENSAL, the notation "alkadienyl" would be **representation-definite** with respect to GENSAL and to the structure represented, i.e., it would be a **proper** and **exact representation**.
>
> Thus, if concrete expressions of GENSAL and of the ECTR are used in practice for representing structures, then the inherent (=factual) extension is to be distinguished from the **extension represented**.

Such cases of representation-indefiniteness may also be conceived and described as cases of **translation-indefiniteness** if the structure to be represented is given by an expression of any structure language, e.g., by "conjugated alkadienyl" in chemists' language. The **proper translation** "alkadienyl" then is **translation-indefinite** *with respect to the language it belongs to* (i.e., GENSAL), *with respect to the expression to be translated* (i.e., "conjugated alkadienyl") and *with respect to the language of the expression to be translated* (i.e., chemists' language).

In the ideal case of an exact representation (or exact translation), the inherent (=factual) extension of the exact representation (or exact translation) is identical with the extension represented (or with the factual extension of the expression translated). In the case of an inexact yet proper representation (or translation), the inherent (=factual) extension of the proper representation (or translation) must strictly include the extension represented (or the factual extension of the expression translated). The inverted case of inclusion or the intersection of these extensions is not permitted (because of the risk of incomplete recall in searches). Thus, in representations (or translations) for the purpose of storage and retrieval of structures, some unavoidable degree of loss of precision is admitted. [To be exact, the conditions discussed for proper representations (or proper translations) are strictly valid only for *extensionally* proper representations (or translations).]

Although representation and translation are closely related to one another in practical use of GENSAL, the two types of expressional definiteness should be distinguished, because a notation may be representation-indefinite in one respect and translation-definite in another. If, in a particular pragmatic situation, a structure is given in chemists' language by the expression "conjugated alkadienyl", and if this structure is to be represented properly in GENSAL and in the ECTR, then "conjugated alkadienyl" must be translated properly into "alkadienyl" in GENSAL, which is properly translated further into the corresponding parameter list of the ECTR. The expression alkadienyl in GENSAL then is representation-in-

definite with respect to the class of conjugated alkadienyl radicals represented, and it is translation-indefinite with respect to "conjugated alkadienyl" of the chemists' language. The paramater list of "alkadienyl" in the ECTR is also representation-indefinite with respect to the structure to be represented, but it is translation-definite with respect to "alkadienyl" in GENSAL.

4.5.3. MEANING-DEFINITENESS AND MEANING-INDEFINITENESS IN CHEMISTS' LANGUAGE

In chemists' language, in as far as it is used as a structure language, the expressions are usually meaning-definite. In some of these expressions, the meaning-definiteness originates in the definitely regulated formalisms of the syntax and semantics of chemists' language. It reflects then, in the same manner as the inherent meaning-definiteness of expressions in GENSAL and in the ECTR, the mere *syntactic* and *semantic* aspects of expressional definiteness. In other expressions, the meaning-definiteness originates in the unrestricted and unregulated versatility of syntactic and semantic means that can be used in chemists' language if it is required in some particular pragmatic situation. This meaning-definiteness then reflects the *pragmatic* aspect of expressional definiteness of chemists' language also.

In some expressions in chemists' language, however, the information given in the intension expressed is not sufficient to determine the structural extension exactly. Such expressions must be seen as **meaning-indefinite** *with respect to chemists' language* in as far as it is used as a *structure* language. This meaning-indefiniteness of expressions in chemists' language may be differentiated into

Vagueness
Uncertainty and **Undecidedness**
Equivocality

These three types of meaning-indefiniteness differ from one another in the peculiar characteristics of the indefinitely determined extension.

4.5.4. VAGUENESS

Examples of **vague expressions** are

(a) "organic group with N, O, S", "organic amine cation"
(b) "electron-withdrawing group", "pharmaceutically acceptable group", "protecting group"
(c) "aliphatic hydrophobic group C6–20", "herbicidal activity imparting group, linked through S, C"

In vague expressions, regarded as abstract, isolated expressions, the indefinitely determined inherent extension is usually a generic partial structure. The bounds of this class, in as far as they are determinable by merely structural information, are vague. Within this area of vagueness, it cannot be decided definitely on solely structural criteria whether a specific partial structure is a member of this class or not. In comparing this class with other structures in automatic or intellectual operations, therefore, a definite decision on matching or nonmatching cannot be achieved for those structures that fall under this area of vagueness.

This area of vagueness, covering specific structures which might possibly belong to the vaguely determined class, or otherwise, is common to the various forms of vagueness:

Sometimes, this area of vagueness covers all conceivable organic partial structures ("organic group").

In other cases, the whole vaguely determined class is an area of vagueness surrounded by a complementary class of specific structures that are definitely not members of the vaguely determined class ("organic group with N, O, S").

In yet other cases, the area of vagueness surrounds, as a border area at the fringes of the vaguely determined class, a hard core of specific structures that definitely are members of the class (herbicidal activity imparting group).

And finally, there may be cases where the area of vagueness lies between such a complementary class and such a hard core (electron-withdrawing group).

The indefinitely determined inherent extension of a vague expression may be determined in a structurally insufficient manner

(a) exclusively by structural information that is not sufficiently exact
(b) exclusively by information on nonstructural properties of the specific partial structures within the class, the nonstructural properties being more or less vaguely connected to structural properties
(c) by the combination of structural information, which may be more or less exact, with nonstructural information, which serves as a restricting qualification

The vaguely expressed inherent intension of a vague expression may be influenced by the context to such a degree that the factual extension (determined by the shifted, factual intension) is different from the inherent extension. In most cases, however, the factual extension remains indefinitely determined, even if it is determined somewhat less indefinitely than the inherent extension. Since this context dependency of the extension of a vague expression usually results again in a vague difference between the inherent and the factual extension, it may be neglected here, as it is neglected in practical use of the GENSAL system.

From the pragmatic point of view of patent applicants or research chemists, vague expressions may be sufficiently definite in their legal or scientific contexts. But they are meaning-indefinite from the pragmatic point of view of an encoder. Vague expressions in chemists' language are therefore encoded in GENSAL as Other Terms, i.e., they are incorporated in this language and in the ECTR as text phrases without any essential syntactic or semantic alterations. Since Other Terms remain the same vague expressions, having the same indefinitely expressed intension and the same indefinitely determined extension in chemists' language and in the GENSAL system, the problem of proper translation is suppressed, but the pragmatic problem of proper comparison of structures with regard to matching-relations remains.

There may be cases of vagueness, produced in manner (c), that seem to be pretty meaning-definite, e.g., "hydrocarbyl group which corresponds to a hydrocarbon boiling between 71 and 93 °C". For this expression, even the exact structural extension could be determined by the intersection of a class of structures with the class of all compounds having a special nonstructural property. Such an exact structural extension, however, is not determined directly by exclusively structural information, but is determinable only indirectly (and individually for each compound checked) in a pragmatically unacceptable and roundabout way through expert knowledge, by connecting structures and nonstructural data via compounds. This case, therefore, is also meaning-indefinite from the pragmatic viewpoint of the encoder. This example shows, however, another pragmatically acceptable way of treating Other Terms.

In parallel with their special treatment as Other Terms (preferably in the modified form of standardized notations within a controlled vocabulary of Other Terms), they can also be integrated into the GENSAL system in the same way as generic nomenclatural terms:

The indefinitely determined extension of nearly all Other Terms is restricted, by restrictive (although usually indefinite) structural or nonstructural information, to a subclass of the class denoted by the generic nomenclatural term "radical"—in the exaggerated example discussed above even to a subclass of "hydrocarbyl". If expressions like "radical" or "hydrocarbyl" are appropriately treated in the GENSAL system (e.g., by user-defined match levels in searching), then Other Terms may be encoded as Other Terms and additionally in parallel, in an appropriate form of syntactic linkage, as generic nomenclatural terms, supplemented by restricting parameter values if possible.

Restrictive structural or nonstructural information in Other Terms, which cannot be represented by these generic nomenclatural terms, can be searched (in the form of the full or the standardized wording of the Other Terms encoded in parallel), and/or it can be used to select relevant answers from the answer set retrieved.

4.5.5. UNCERTAINTY AND UNDECIDEDNESS

Among inherently meaning-indefinite expressions of chemists' language that are, as abstract notations, **uncertain expressions** and that are, as concrete expressions in a particular pragmatic context, **decided expressions** or **undecided expressions** one may distinguish between

(a) singly uncertain expressions: "phenyl substituted by Cl", "phenyl substituted by alkyl".

(b) multiply uncertain expressions: "aryl", "aliphatic group".

The inherent extension of the notation "phenyl substituted by Cl", i.e., the extension of the abstract notation, regarded in isolation from any context, is indefinitely determined: it is uncertain whether this inherent extension is identical with the extension of "phenyl substituted by one Cl only", or with that of "phenyl substituted by 1–5 Cl". The indefinitely determined inherent extension of the **uncertain expression** "phenyl substituted by Cl" may be conceived as a set of two definite extensions that is determined, by an expressed **uncertain alternative**, to be a set of two alternatively possible extensions of one notation. The narrower class is a finite class containing three members (*o*-, *m*-, and *p*-chlorophenyl); the broader extension is a finite class containing 19 members. The narrower extension is included by the broader extension.

The inherent extension of such a singly uncertain expression is indefinitely determined in a way that is clearly different from that of the indefinitely determined inherent extension of a vague expression. The vaguely expressed intension of a vague notation determines, in a vague manner, an indefinite inherent extension the characteristic of which is an indefinite area of vagueness; the uncertainly expressed intension of the singly uncertain notation gives mainly definite structural information, but it contains also an uncertain dual alternative of structural information which determines, in an uncertain manner, an indefinite inherent extension being a set of two definite, alternatively possible extensions. The dual alternative concerns one **uncertain structural feature**, which is the frequency of substitution at a ring in the example discussed above. The alternatives are the counts 1 or 1–5.

In particular pragmatic situations, however, when the intensions may be influenced by the context, the *factual* extension of the concrete expression "phenyl substituted by Cl" may be a definitely determined extension, although the inherent extension of the abstract expression is an indefinitely determined extension. If this concrete expression is, for example, a part of a complex Markush formula containing the final phrase "...with the proviso that the molecule does not contain more than one halogen atom", or if this expression, occurring in the main claim of a patent, is instantiated in an example by "2,4,6-trichlorophenyl", then from the pragmatic viewpoint of an encoder who has to encode these concrete expressions into GENSAL, the inherently uncertain alternative is factually concretized as a **decided alternative**, which has been decided in favor of the narrower or broader extension respectively. Thus the uncertain abstract notation "phenyl substituted by Cl" may be concretized in two different contexts as two different concrete expressions, which are factually **decided expressions** with different factual extensions, although their isolated notations are identical in chemists' language. For uncertain expressions in chemists' language, therefore, the factual extension is not context-independent. The **inherent meaning-indefiniteness** of an expression in chemists' language, in this case the **uncertainty**, may be concretized in particular contexts as a **factual meaning-definiteness**, in this case as **decidedness**. The expression then is uncertain, yet decided.

In other pragmatic contexts in patents, on the other hand, the uncertain expression "phenyl substituted by Cl" may be instantiated in examples by mono-*p*-chlorophenyl alone, which gives no decisive information with regard to the uncertain alternative. The inherently uncertain alternative then is concretized factually as an **undecided alternative**, and the uncertain expression is concretized as an **undecided expression**. In this case, the factual extension remains identical with the inherent extension, and it remains an indefinitely determined extension.

The example "phenyl substituted by alkyl" is to be conceived quite analogously; in this case, too, the indefinitely determined inherent extension of this uncertain expression is a set of two alternatively possible definite extensions, being determined by one uncertain alternative concerning one uncertain structural feature. The sole difference is that the two definite extensions are two potentially infinite classes, which are nevertheless different extensions, the narrower being included by the broader.

The pragmatic problems posed by such singly uncertain expressions for translation into GENSAL and for retrieval by the GENSAL system are obvious:

Encoding the undecided concrete expression "phenyl substituted by Cl" simply in the same manner as the meaning-definite "monochlorophenyl" would be pragmatically incorrect. The intensional difference and the difference in relevance between the two expressions could not be distinguished in a search for monochlorophenyl; and in a search for dichlorophenyl, the undecided expression would not be retrieved. This could be regarded as a case of incomplete recall.

Encoding the above undecided expression simply in the same manner as "phenyl substituted by 1–5 Cl", just to be on the safe side, would also be pragmatically inadequate. The intensional difference and the difference in relevance between the two expressions could not be distinguished in a search for pentachlorophenyl.

Encoding the two concrete expressions "phenyl substituted by Cl", which are decided in favor of the narrower or broader extension in the same manner as the corresponding expressions "monochlorophenyl" or "phenyl substituted by 1–5 Cl", respectively, would be pragmatically acceptable. There is only a negligibly small intensional difference between an inherently meaning-definite expression and an uncertain yet decided expression having the same factual extension.

The pragmatic problems are even more intricate for multiply uncertain expressions like "aryl":

In some patents, the context of "aryl" suggests that the factual extension is to be seen as restricted to unsubstituted carboaryl; in other patents, the context

clearly determines the factual extension as including possible heteroaromaticity and/or possible substitution on aromatic rings.

For "aryl", the indefinitely determined inherent extension is not determined by a single uncertain dual alternative, but by several uncertain dual alternatives concerning several uncertain structural features, such as heteroaromaticity, fusion with nonaromatic rings, and different kinds of substitution on the ring(s) by alkyl, by other rings, by heteroconnected groups, etc. The uncertain dual alternative for each of these uncertain features is the alternative between strict exclusion of this feature from the extension of "aryl" or possible inclusion in this extension, i.e., the uncertain alternative for each of these features is the alternative between obligatory absence or possible presence of this feature in specific structures of the class given by "aryl".

Each uncertain alternative determines two extensions: a narrower extension by obligatory absence and a broader extension by possible presence of this feature. Both extensions are potentially infinite classes. The narrower extension, again, is included by the broader extension.

Provided that a set of n uncertain features for "aryl" is defined clearly by pragmatic conventions, then the indefinitely determined inherent extension of the uncertain expression "aryl" is a set of 2^n different definite extensions, and it is uncertain which one of these extensions is really meant by an author using "aryl". This "cluster" of 2^n distinct extensions is constituted by combining each of the two extensions resulting from each uncertain alternative with each pair of extensions of all other uncertain alternatives.

The union of all those extensions, which are produced by the possible presence of all uncertain features, is the **broadest conceivable extension** of "aryl".

The intersection of all those extensions produced by the obligatory absence of all uncertain features is the **narrowest conceivable extension** of "aryl", which would reasonably be identical with the extension of "unsubstituted, exclusively benzenoid carboaryl".

In a particular pragmatic context, none or usually only some of the uncertain alternatives will be decided alternatives, determining the factual extension of the uncertain expression "aryl" in this context. Decidedness and undecidedness then are particularized into decided and undecided features, which themselves are represented by decided and undecided parameter values.

The problems of an adequate pragmatic treatment of such multiply uncertain expressions in storage and retrieval are, in principle, analogous to those discussed with regard to singly uncertain expressions like "phenyl substituted by Cl", although they are more complex. The problems may be raised only by pointing to the evident inadequacy of both of these simplifying ways of defining the parameter list of "aryl": either in terms such that it denotes simply either the narrowest conceivable extension or the broadest conceivable extension. These problems will be the subject of another publication.

4.5.6. EQUIVOCALITY

Equivocal expressions, quite common in natural languages, are disapproved of in scientific languages. Nevertheless, they sometimes occur as meaning-indefinite expressions in chemists' language. For the simple partial expression sulfonylamino, for example, as it occurs in Markush formulas, it may not be quite clear in some pragmatic contexts whether the doubly connected group is connected as $-SO_2NH-$ or as $-NHSO_2-$.

In chemists' *structure* language, equivocal expressions are usually *ambiguous* in a literal sense, i.e., they have only *two* meanings. Equivocality is not restricted to expressions for doubly connected specific partial structures, it also occurs in expressions for singly connected and generic partial structures (e.g., "aralkyl", "alkaryl", "arylalkyl", and "alkylaryl"). In such composite generic expressions too, equivocality is caused by the possible inversion of connectedness.

In some cases, the inherent equivocality of an abstract expression may almost undoubtedly be resolved by the context of the concrete expression, e.g., in "R = sulfonylamino, -CONH-, -COO-, or $-SO_2O-$". But in other cases, equivocality may be just caused by the syntactic context, e.g., by the pictorial arrangement of signs within a formula or by the order of signs within a line, e.g., in "$P(arylalkyl)_3$". The example "arylalkyl", by the same token, shows that expressions may be equivocal and uncertain at the same time.

Equivocality in its pure form is clearly different from **uncertainty**, even from the uncertainty of singly uncertain expressions. In both cases, for an exclusively equivocal expression like "sulfonylamino" and for a singly uncertain expression like "phenyl substituted by Cl", a set of two alternatively possible extensions is indefinitely determined by an uncertain dual alternative. But for an uncertain expression, the narrower extension is included by the broader, and for an equivocal expression, both extensions are completely different in the sense that they are differently connected (and therefore different) specific partial structures or that they are classes which have no specific partial structure in common, i.e., which do not intersect.

If equivocality can not be resolved by the context, both alternatively possible extensions of the equivocal expression should be represented as alternative partial structures by the translationally corresponding expressions of GENSAL.

5. STRUCTURAL EXPRESSIONS IN THE GENSAL SYSTEM

5.1. PROPERTIES OF STRUCTURAL EXPRESSIONS

Particular properties of structural expressions may be described by characterizations like "specific", "partial", "simple", "meaning-definite", "vague", etc., or by characterizations like a "structure diagram", "line formula", "standard text term", "generic nomenclatural term", etc. These characterizations have been used as terms in the previous sections in a rather heuristic manner, mostly in connection with exemplifications taken from chemists' language. They can be defined and used as terms in an exact manner, however, in regard to their application to the formalized structure languages used in the GENSAL system. Their application to expressions in chemists' language is also definite enough, in most cases, by analogy with the translational correspondence between expressions in GENSAL and chemists' language. These terms may be used as terminological components of a theoretical metalanguage in order to discuss pragmatic problems in the use of structure languages. In particular, they may be used for making typological differentiations (e.g., "specific partial expression" or "generic full expression").

The differentiation of properties described in a terminologically differentiating manner may refer:

- to merely syntactic properties of structural expressions (as exemplified by the differentiation between a *standard text term* and a *structure diagram*)
- to semantic properties of particular syntactic forms (e.g., the differentiation between *specific* and *generic nomenclatural terms*)
- to pragmatic properties of particular notations used in different concrete pragmatic contexts (e.g., the differentiation between two concrete expressions in

GENSAL having the same notation "alkadienyl"; the one being *representation-indefinite* in its pragmatic context, the other being *representation-definite* in another pragmatic context).

As these examples show, properties of structural expressions are not only of theoretical interest. Different properties necessitate differing treatment of expressions in practical operations, e.g., in the steps of translating from chemists' language to GENSAL and the ECTR, to reduced graphs or to fragment screens, in search operations, and in the assessment of retrieved answers. Especially in automatic search operations, particular properties are of high relevance to the sequence of steps in a search algorithm:

A particular expression can be treated properly in accordance with its particular properties only if these properties have been recognized by the algorithm in an earlier step.

A node in a reduced graph is derived from a particular (simple or composite) expression in the ECTR. In a two-stage search at these two levels of representation [i.e., the first stage comparing nodes of the reduced graphs (of file and query structures) and the second stage, the refined search, comparing the corresponding parts of the ECTRs], the whole process can be facilitated and shortened if the **derivation-type** of each node (i.e., the particular property of being derived from a particular type of expression in the ECTR) is known at the time of operations on the nodes. This is discussed further below (in Section 5.7).

Operations on expressions, therefore, can be facilitated and speeded up if the expressions have been marked beforehand with labels designating their operationally relevant properties in a concise form. This is the basis of the concept of using determinant-values.

5.2. THE CONCEPT OF EXPRESSION-DETERMINANTS

The operationally relevant properties of an expression can be described by a bit string; each position in this bit string stands for the potentiality of a particular property. The attribution of this property to the expression is designated by the bit value "1", the nonattribution by the bit value "0". Each particular position in this bit string stands, therefore, for a particular **expression-determinant**, having one of two mutually exclusive **determinant-values**. The two determinant-values, which are denoted within the bit string by a concise standardized notation, may also be denoted in the theoretical language by other notations which are assigned to expressions as adjectival or substantival terms, e.g., "meaning-definite"—"meaning-indefinite" or "meaning-definiteness—"meaning-indefiniteness". The abstract expression-determinant itself, if the terminological need exists, can be designated by a single term, e.g., by the artificial term "meaning-definity".

The different expression-determinants of the bit string may be defined according to operational needs, and the bit string denoting the respective determinant-values may be associated with simple or composite expressions within the internal GENSAL representation, within ECTRs, and within reduced graphs in a syntactically and operationally suitable manner. A precondition for the accelerating effect of this use of determinant-values in search operations is that they can be derived algorithmically prior to the operation that needs to be accelerated. They may thus be derived in the input process, creating internal representations for permanent storage, or they may be derived in an intermediate process during the search, creating only temporary representations of those structures that have already been selected, e.g., as a result of a fragment search, as candidates for matching.[1,14]

From the technical point of view, this use of bit strings may seem to be similar to the use of bit strings in fragment searching. From the theoretical point of view, however, there are significant differences:

In fragment screening, the twin bit string (i.e., the combination of the MUST bit string with the union of the MUST and MAY descriptors) may be seen as the composite expression of a generic full structure in a special structure language. This language of fragment screens can be processed rapidly, but only at the expense of the expressive power of its expressions, which are highly representation-indefinite. The combination of the two corresponding bits in the twin bit string designates the obligatory absence, the obligatory presence, or the possible presence of a particular structural feature in fragments of a structure. In fragment screening, therefore, bits are used to denote *structural* properties of *structures*, i.e., of nonlinguistic objects.

In a bit string used for denoting the values of expression-determinants, however, bits are used to denote *nonstructural* properties of a structural *expression*, i.e., of a linguistic object. (These properties, in as far as they are properties of *structural* expressions, may have an indirect reference to the structural properties of the structures so denoted, but they are nonetheless not structural but syntactic, semantic, or pragmatic properties of *expressions*.)

Bit strings denoting the values of expression-determinants can be incorporated into formalized structure languages to facilitate operations on expressions. They thus become parts of structural expressions, but they *are not* structural expressions themselves. These special parts of notations may be conceived as expressing *explicitly* that part of the intension that would otherwise be expressed *implicitly* only, as operational information, by the notation and its pragmatic context.

In principle, any arbitrary distinct property of structural expressions can be used as one of two mutually exclusive determinant-values, the other being simply the absence of this property. Not infrequently, however, two properties of structural expressions are conceived and described (e.g., by the terms "specific"–"generic") as forming pairs of mutually exclusive properties. In such cases, these pairs of properties can be defined, with respect to the expressions of a particular formalized structure language, in such a manner that the terms denoting the two properties exclude one another by mutual negation and are applicable to every expression of this language. Such pairs of properties thus can be conceived and used as the two determinant-values of one expression-determinant.

5.3. DETERMINANT-VALUES OF EXPRESSIONS IN GENSAL AND IN THE ECTR

Several operationally relevant properties, which may be used as the values of expression-determinants, have already been introduced implicitly in terms which have been used primarily for typological purposes and in a heuristic manner. These determinant-values will now be redefined in relation to GENSAL and the ECTR. To simplify matters, the more familiar expressions of chemists' language are used in examples instead of the translationally equivalent GENSAL and ECTR notations. The expression-determinants and their values are not defined stringently and completely here. Rather, it will be shown that they can be defined exactly in an operational sense: for any expression in GENSAL or in an ECTR, it is definitely decidable, by an algorithmic procedure based solely on syntactic criteria, whether one or the other determinant-value of a particular expression-determinant is to be attributed to this expression.

The determinant-values denoted by the terms **partial** and **full** can obviously be derived simply from the syntactic form of expressions in GENSAL and the ECTR. By analogy, the properties denoted by these terms may also be attributed to expressions in chemists' language.

The determinant-values denoted by the terms **"simple"** and **"composite"** too can be derived from the syntax of the ECTR and from the syntactically definite rules governing the translation from GENSAL into the ECTR:

An ECTR expression is simple if it is a distinct parameter list or a distinct connection table (which, in most cases, is a *partial* connection table, denoting a specific *partial* structure).

All other ECTR expressions containing two or more such syntactic units are composite expressions.

Simple (or composite) GENSAL expressions and expressions in chemists' language are those which are translated into simple (or composite) expressions in the ECTR.

The determinant-values denoted by the terms **homogeneous** and **inhomogeneous** can be defined and denoted in a similar manner, by reference to exclusively syntactic criteria:

In the ECTR, all simple expressions are homogeneous, and all composite expressions that comprise exclusively connection tables or exclusively parameter lists are also homogeneous [e.g., the expressions which correspond translationally to "phenyl", "cycloalkyl", "phenyl substituted by one or more ethyl", "R_1-CH=CH-R_2; R_1 = methyl or ethyl"; "R_2 = phenyl or cyclohexyl", "cycloalkyl substituted by one or more alkyl", "-alkenylene-(C2-4)-alkyl(C1-5)"].

Inhomogeneous expressions in the ECTR are all those composite expressions which are combinations of connection table(s) with parameter list(s) [e.g., the translations of "phenyl substituted by one or more alkyl", "R_1-CH=CH-R_2; R_1 = alkyl(C1-5)"; "R_2 = phenyl or cyclohexyl", "cycloalkyl substituted by one or more ethyl", "-alkenylene(C2-4)-*t*-butyl"].

Homogeneous (or inhomogeneous) expressions of GENSAL and of chemists' language are those which are translated into a homogeneous (or inhomogeneous) expression in the ECTR.

Another pair of determinant-values, denoted by the terms **"segmented"** and **"nonsegmented"**, must be distinguished carefully from the determinant-values denoted by "homogeneous" and "inhomogeneous" and "simple" and "composite". The division into distinct specific partial expressions is somewhat arbitrary in composite expressions that are homogeneously composed exclusively of specific partial expressions. The extension denoted by the expression "phenyl substituted in the para position by chloromethyl, ethyl, or *n*-butyl" could equally well be denoted by the enumeration of three structure diagrams or by the expression "*p*-C_6H_4-CH_2-R; R = Cl, methyl, or *n*-propyl", which is extensionally identical yet differently divided. Ultimately, every *generic* full expression that is homogeneously composed from exclusively specific partial structures can be resolved into a finite set of *specific* full expressions. In contrast to this changeable linkage between specific partial expressions, the linkage between expressions like "alkyl" and other partial expressions cannot simply be shifted or avoided completely; it is a rigid cutpoint of **segmentation**. If the segmentation is different in query and file structures, then it may cause particular problems in search operations, e.g., in comparing pairs selected from the following expressions: "-CH=CH-alkyl(C6-8)", "-CH=CH-CH-(CH_3)-(CH_2)$_2$-alkyl(C2-4)", "-alkenylene(C3-6)-*t*-butyl", "-alkenylene(C7-9)-alkyl(C2-5)".

In the ECTR, composite expressions are segmented if they contain one or more segmentation cutpoints between two parameter lists, or between a parameter list and a connection table.

In the ECTR, simple expressions are nonsegmented, as well as all those composite expressions which are exclusively composed of distinct connection tables.

Segmented (or nonsegmented) expressions in GENSAL and chemists' language are those which are translated into segmented (or nonsegmented) expressions in the ECTR.

Semantic criteria must seemingly be used in defining and deriving the determinant-values denoted by the terms **"generic"** and **"specific"**: a specific expression denotes a specific structure; a generic expression denotes a generic structure, i.e., a class of specific structures. However, these determinant-values too can be derived by exclusively syntactic criteria in GENSAL and in ECTR, because classes of structures can be denoted in these languages only by special types of syntactic formalisms. This will be described in Section 5.5.

Because the values of an expression-determinant can be derived strictly on syntactic grounds alone, they can be conceived as values of a *one*-valued *singulary* function; the range of arguments of this function consists of all conceivable expressions of a particular structure language; the range of values comprises two values only. Application of this function to any *single* argument yields only *one* of two mutually exclusive values.

This strict concept of the syntactic derivability and of the use of determinant-values, which are properties of structural expressions, is possible for other semantic properties of expressions in GENSAL and in the ECTR too, in as far as these properties are strictly *semantic* properties of *abstract* expressions. The determinant-values denoted by the terms **inherently meaning-definite** and **inherently meaning-indefinite**, for instance, are derivable from the purely syntactic form of expressions in GENSAL: only Other Terms (which are marked syntactically by inclusion in single quotes) and composite expressions containing Other Terms are inherently meaning-indefinite; all other terms are inherently meaning-definite in the current version of GENSAL. (If the particular uncertainty and undecidedness of expressions in chemists' language were to be treated in future versions of GENSAL by adequate syntactic means, then such uncertain expressions would be inherently meaning-indefinite too.)

Those properties of *concrete* expressions in the GENSAL system which appear to be semantic properties but which are really *pragmatic* properties in as far as they are influenced by the pragmatic context cannot be handled in this strict concept of expression-determinants whose values are derivable by exclusively syntactic criteria. The property of **representation-indefiniteness** attributed to a particular concrete "alkadienyl" in GENSAL and to the translationally corresponding parameter list in the ECTR, for example, cannot be recognized automatically. But it could be handled as a determinant-value in a less strict manner; it could be encoded manually, after an intellectual assessment of the pragmatic situation, in a suitable syntactic form, e.g., by a special mark. (The same would hold for the pragmatic properties of **factual meaning-definiteness** and **factual meaning-indefiniteness**, and for those special types of these properties that are called **decidedness** and **undecidedness**, respectively.)

5.4. THE CONCEPT OF PARAMETER-DETERMINANTS

The concept of determinants and determinant-values that has been applied to simple and composite expressions in the previous section may be applied to parts of a particular type

of simple expressions too, viz., to those distinct parts of a parameter list which denote the distinct parameter values. Within the parameter list of alkadienyl, e.g., the syntactic unit "E⟨2⟩" denotes a particular status of a *structural feature* in the structure alkadienyl, viz., the obligatory presence of two ethylenic double bonds, but it does not denote a *structure*, and it is therefore not a structural expression itself. Special semantic properties of these syntactic units—e.g., the property of expressing the obligatory presence, the obligatory absence, or the possible presence of the respective structural feature—can be treated as determinant-values. These determinant-values may be distinguished as values of **parameter-determinants** from the values of the **expression-determinants** discussed in the previous section.

The values of *parameter*-determinants can be arranged as bit strings in such a manner that these bit strings can be used as **Determinant-Screens** in rapid preliminary screening operations similar to fragment screening, for comparing (genuine generic or derived specific) parameter lists with one another. This concept, used as the concept of **Reduced Parameter Values**, will be discussed in more detail in the following paper in this issue.

5.5. TYPES OF CLASS-CONSTITUTING MECHANISMS

In the use of structure languages, the concrete generic expressions of generic (full or partial) structures **denote** an **inherent extension**, which is a class of two or more specific structures, and they **represent**, more or less exactly, a **represented structure** (i.e., the structure to be represented in the given concrete pragmatic situation), which is also a class of specific structures.

Bit strings used for fragment screening and reduced graphs determine the inherent extension by a generalizing form of information, viz., by an intension using reduced structural concepts which are not sufficient for representing structures exactly and which, therefore, are not usual in "normal" structure chemistry. The expressions of these structure languages are highly representation-indefinite, but they are useful for rapid preliminary search operations.

In structure languages with high expressive power (chemists' language, GENSAL, and the ECTR), generic expressions not only represent the structure to be represented in a sufficiently exact way, they also express the conceptual mechanism by which the generic structure is constituted as a class of specific structures by means of the usual structural concepts. The intension expressed by the notation contains not only the information about structural "modules" (i.e., about atomic fragments or multi-atom specific partial structures, or about other characterizing values of structural features, e.g., of aromaticity and other ring features, as in "nonaromatic fused carbocycle"), it also contains the instructions for constructing every single specific structure of the class by putting together the modules in varying combinations. As will become clearer in the following examples, this instruction is given explicitly by purely syntactic morphology (i.e., by the graphical or typographic texture) of the notation, and/or it is given implicitly by the inherent meaning of the notation, e.g., of a generic nomenclatural term.

Four different mechanisms of class-constitution can be distinguished, although two or more of these mechanisms may coexist in one generic (full or partial) expression. These mechanisms differ from one another by the way in which the combinations of modules are varied. They may be conceived, therefore, as different principles (or types) of variation. These types of variation can be defined in formalized structure languages with reference to exclusively syntactic criteria. In the following, they are exemplified only by expressions from chemists' language, illustrating the pure and isolated types of variation:

p-variation (position-variation) of the possible mutual *positions* of partial structures. Examples: "monochlorophenyl", "monochlorophenol", or the corresponding structure diagrams indicating the variable position of substitution on the ring by the usual graphic means. Evidently, at least one of the partial structures combined by p-variation must be a *specific* partial structure with two or more possible points of attachment. (An imaginary variation of the mutual position in "monochloroaryl" or in "monoalkylaryl", for example, would not contribute to the constitution of the denoted class because the generic partial structures denoted by "alkyl" and "aryl" are conceived, per se, as structures with varying attachment.)

s-variation (substituent-variation) of the possible *substituents* at a fixed position of a partial structure. Examples: "phenyl substituted in the para position by F, Cl, or Br"; "phenol substituted in the para position by F, Cl, or Br". Since a generic nomenclatural term is to be conceived as denoting one structure, being one substituent, the expressions "phenyl substituted in the para position by halogen" and "phenyl substituted in the para position by C1–3 alkyl" are not examples of s-variation but of another type of variation (viz., of h-variation).

f-variation (frequency-variation) of the possible *frequencies* of partial structures. Examples: "phenyl substituted in the para position by $-(CH_2)_n$-Cl; $n = 1–3$". In nonpolymeric structures, the upper limit of the frequencies must be finite.

h-variation (homology-variation) of the possible combinations of structural features within the limits determined by the intension of a (simple or composite) expression like "cycloalkyl" (or "alkane", "alkyl", "phenyl substituted in the para position by C1–3 alkyl"), which denotes a class other than by p-, f-, or s-variation. Thus, h-variation is the variation within a (finite or infinite) series of specific structures which are *homologous* to one another in a very wide sense of *homology*; the series is determined by the concept of a serial order of specific structures which are similar in regard to particular criteria and which are different with respect to certain other particular criteria. This h-variation is the usual class-constituting mechanism of simple generic nomenclatural terms, and the series is usually constituted by variation with respect to three or more criteria. In less frequent expressions, which may be combinations of generic nomenclatural terms with qualifying restrictions or which occasionally may not show any resemblance to generic nomenclature at all, the series may be constituted by variation with respect to one or two of these criteria, e.g., with respect only to the following:

number of atoms within a chain (e.g., "*n*-alkyl", "*n*-alkyl C1–8")

branching of a chain (e.g., "primary dodecanol")

branching and point of attachment of a chain (e.g., "any dodecyl radical")

position of multiple bonds (e.g., "*n*-dodecenyl", "*n*-dodecadienyl")

number, kind, and position of multiple bonds (e.g., "unsaturated nonbranched C8-hydrocarbyl chain")

ring size (e.g., "monocycloalkyl", "$-NR_1R_2$ with $R_1 + R_2$ = *n*-alkylene, forming a saturated ring")

pattern of ring fusion and point of attachment (e.g., "fused carboaryl consisting of five benzene rings")

number of rings (e.g., "aromatic hydrocarbon consisting of benzene rings fused in a linear order")

point of attachment and permutation of heteroring atoms (e.g., "1N, 1O, and 1S containing, 6-membered, saturated heterocycle, connected by C or N")

As several of the examples of h-variation show (e.g., "any dodecyl radical", as compared to "alkyl"), special series of isomers which are merely subsets of the usual homologous series must be conceived as being constituted by h-variation.

Furthermore, h-variation should be understood too in the wider sense of homology, as a variation with respect to the membership of a group or a period or an otherwise defined class within the periodic system, e.g., "halogen", "heteroatom", "transition metal", "alkaline metal", or "lanthanide".

In principle, obviously, all homologous series in this broad sense can be represented in a sufficiently definite manner by parameter lists—the expressive power of an enlarged set of parameters only needs to be adjusted to this requirement, as has been done in the work at IDC.[6] For most of these series, the use of a parameter list is the only practicable means of representation; whereas "any butyl radical" could still be encoded by resolving the expression into the four specific structure diagrams, such a resolution is not practicable for "any dodecyl radical" (which can be encoded most conveniently as "alkyl ⟨12⟩"). For other finite series, the use of a parameter list is an adequate means of expressing an intensional difference, e.g., that between "halogen" and "F, Cl, Br, or I" (cf. Section 3.3).

5.6. VARIANCE-VALUES

The occurrence or nonoccurrence of each of these four types of variation in a generic expression may be conceived as a particular property of this expression, and these properties can be treated as values of four expression-determinants:

The two mutually exclusive values of the determinant called **p-variance** (position-variance) are denoted by the terms **"p-variant"** and **"non-p-variant"**.

Analogously:

s-variance (substituent-variance): **s-variant** and **non-s-variant**

f-variance (frequency-variance): **f-variant** and **non-f-variant**

h-variance (homology-variance): **h-variant** and **non-h-variant**

For expressions in GENSAL and the ECTR, these **variance-values** are derivable purely from the syntactic form. They may be denoted and used, together with the operationally relevant values of other expression-determinants, by using an appropriately defined bit string. Thus, the GENSAL or ECTR expression which corresponds translationally to the expression "phenyl substituted by one or more Cl and/or alkyl", can be characterized by the following determinant-values: p-, s-, f-, and h-variant, generic, partial, composite, inhomogeneous, segmented, and meaning-definite.

The four different variance-values attributed to a particular expression are mutually independent of one another. The values of other expression-determinants, however, are not completely independent of one another nor of the variance-values. Some interdependencies are as follows:

An expression which is inhomogeneous and/or segmented and/or p-variant and/or s-variant must be a composite expression.

No specific expression is p- or s- or f- or h-variant.

A generic expression is p- and/or s- and/or f- and/or h-variant.

The determinant-values denoted by the terms "specific" and "generic" are thus not only derivable from the variance-values, they could also be replaced by the variance-values. This elimination of redundancy is useful in computer operations, but the terms specific and generic are to be preferred in theoretical discussions.

For talking about structures, the values of expression-determinants may be attributed not only to *expressions* but also to the *structures* represented by these expressions.

In principle, every generic (partial) structure being p-, s-, or f-variant may be resolved into a finite set of *alternative* (partial) structures, which then are non-p-, non-s-, and non-f-variant, but which may still be h-variant. Since this issue is of relevance to special search operations (as discussed in the following paper in this issue), it is useful to condense the values of p-, s-, and f-variance to the values of a derived determinant called **a-variance** (alternativity-variance), where a structure is **non-a-variant** if it is non-p- and non-s- and non-f-variant, otherwise it is **a-variant**.

If Other Terms are integrated into the GENSAL system in the same way as generic nomenclatural terms, as discussed in Section 4.5.4, then they may be conceived as meaning-indefinite h-variant expressions. If they are treated differently, as in the current version of GENSAL, then they may be conceived as meaning-indefinite and **v-variant**; i.e., the class denoted indefinitely by a vague expression is conceived as being constituted by a fifth type of variation, viz., by **v-variation** (vagueness-variation), which is a special imprecise mechanism which cannot determine a class exactly.

5.7. APPLICATION OF EXPRESSION-DETERMINANTS TO REDUCED GRAPHS

The nodes of a reduced graph are expressions in a structure language with low expressive power. By translating an ECTR into a reduced graph, particular parts of the ECTR—or even the whole ECTR of a generic full structure—"collapse" to one node of the reduced graph. Such parts of the reduced graph may be simple or composite expressions, or they may even be (or may contain) only a part of a simple expression. The connection table of *p*-ethylphenyl, for instance, is a simple expression in the ECTR. Nevertheless, two different nodes of a cyclic/acyclic reduced graph are derived from the cyclic and acyclic parts of this one connection table.

The special expression-determinants and determinant-values that are defined for application to ECTR or GENSAL expressions thus cannot be applied simply and in the same manner to the nodes of a reduced graph. For application to nodes, however, other special expression-determinants and their values may be defined with respect to the reduced graph language by considering the translational correspondence between a node and that part of the ECTR from which the node is derived. Because of this translational correspondence too, these determinants and values for nodes may be denoted, to simplify matters, by using the same terms as are used to characterize expressions in the ECTR and GENSAL. A node then can be referred to as h-variant if it is derived from a part of ECTR which is, within the boundaries of this node, an h-variant expression.

All other terms denoting the values of expression-determinants with respect to the ECTR or GENSAL can be redefined and reused for application to nodes in the same manner. Thus, the notation h-variant, when assigned to an expression in the ECTR, denotes the property of *being* h-variant, but the same notation, when assigned to a node of a reduced graph, denotes the property of being *derived* from an h-variant part of the ECTR. A bit string denoting the values of the expression-determinants for a node therefore may be

conceived as denoting the **derivation-type** of this node.

5.8. FILE LANGUAGES AND QUERY LANGUAGES

During a particular search involving operations at several different levels of representation (e.g., fragment, reduced graph, and ECTR levels), a particular **query structure**—which may be a specific structure or a generic structure—is to be compared with all full **file structures** stored in the database. This comparison is regulated by the special search purpose(s) of this particular search, e.g., by the requirement for distinct match levels in distinct parts of the query structure, or by defining free sites, open for further substitution, for particular parts of the query structure. These special search purposes, optionally particularized for distinct parts of the query structure, are to be conceived and can be implemented in terms of special matching-relations, e.g., in terms of class relations or of the relation of substructural embedment (cf. Section 4.3).

The comparison of structures is performed by examining the expressions that represent the query structure and the file structures more or less exactly at the different levels of representation. At a particular level of representation, for evident reasons of technical feasibility, the syntax and semantics of **query expressions** should be as similar as possible to the syntax and semantics of **file expressions**. In this respect, the whole system of different **file languages** at different levels of representation—including the translational and representational interrelations between these file languages and the meaning-relations within each of these file languages—could be seen as having a mirror-symmetric counterpart formed by the analogous system of **query languages**. The plane of symmetry then would symbolize the comparison process in search operations. In this respect, there seems to be almost no difference between a file language and the corresponding query language at the same level of representation.

The expression of a query language, however, has to express not only the mere structural information by which the denoted query structure is determined, it also has to express, in addition, that part of the intension which is operational information about the appropriate treatment of this query expression in search operations with respect to the special matching-relation(s) required. In contrast to file expressions, therefore, query expressions must contain additional syntactic means for expressing special, optionally particularized search purposes. In this respect, query and file languages are partly asymmetric counterparts.

The syntactic means of expressing particular search purposes in a query language are to some extent arbitrary, provided they are algorithmically processable and enable the user to differentiate and to particularize the search purposes as finely as is required for pragmatic reasons. The examples discussed in the following, therefore, only show *possible* notational implementations of these differentiations of search purposes in GENSAL. They could equally well be replaced arbitrarily by other implementations instead. The simple examples furthermore illustrate merely some possible differentiations of search purposes with regard to only one kind of differentiation, viz., to the differentiation of user-defined match levels. The file expressions in these examples are formulated on the assumption that GENSAL contains "*t*-alkyl" as a standard text term. (Since surface GENSAL, which is primarily an *input* language, needs to be stored permanently to serve as a *display* language, it can be conceived as a *file* language.) Query expressions are tentatively proposed for the corresponding surface query language of the GENSAL system.

(a) For the partial structure *t*-butyl, as a part of a full query structure, the user might wish to restrict the answer set retrieved to those partial file structures which are described specifically by the file expression "*t*-butyl", by the corresponding structure diagram, or by the corresponding line formula "$(CH_3)_3C$". The query expression for this search purpose could be simply "*t*-butyl", the corresponding structure diagram, or the line formula. The matching-relation required by the user in this case is that of (extensional and intensional) identity between query and file partial structure.

(b) The user, however, might wish to get in a structure search different answer sets with different levels of *relevance* in terms of different levels of *specificity*. He might wish then to divide the search into separate searches [e.g., into one search according to (a)], retrieving exclusively hits in which *t*-butyl is specifically described and another search retrieving exclusively those structural generalizations of *t*-butyl which are denoted by h-variant file expressions such as "*t*-alkyl ⟨C4–8⟩", "*t*-alkyl ⟨4–20⟩", or "*t*-alkyl", provided that the level of "*t*-alkyl" is not exceeded (e.g., by such generalizations as "alkyl", "hydrocarbyl", or "radical"). This search purpose could be expressed by "*t*-butyl TO *t*-alkyl". (Any other delimiter, e.g., "UP TO", "GEN", or "LIMIT", or a special syntactic form, e.g., "*t*-alkyl\4\", could be introduced into the surface query language of the GENSAL system.) The special matching-relation (between the query structure and the file structures required) that is required by the user in this case, and its technical implementation, will be discussed in more detail below in this section and in the following paper in this issue.

(c) If the user wishes to combine the two different search purposes (a) and (b) in one search retrieving one answer set, then the query expression could be "*t*-butyl/*t*-butyl TO *t*-alkyl".

(d) If the user does not wish to define any upper level of generalization in a search analogous to (b), then the query expression could be "*t*-butyl TO radical" or "*t*-butyl NOLIMIT".

(e) If the user wishes to get exclusively those specifically described specific partial structures which are included by the class denoted by "*t*-alkyl ⟨4–8⟩", then the query expression could be, for example, "*t*-alkyl! ⟨4–8⟩" or "SPEC *t*-alkyl ⟨4–8⟩".

Further, even finer differentiations of search purpose could be expressed and operated similarly by appropriate syntactic means. If a retrieval system were to mix both specific full and generic full structures with one another (either physically in a unified file or virtually from the viewpoint of the user who gets a single answer set from one search in separate files), then, in a search for specifically described structures, a differentiation could be made between genuine specific full structures and generic full structures matching merely by specific matching-paths through these generic structures (cf. Section 4.3).

These examples of possible notations in the surface query language indicate that translationally corresponding measures must be taken in the syntax of the query ECTR too. Evidently, the use of appropriately defined determinant-values is useful for controlling the operations necessary for the comparison of query and file expressions with regard to special search purposes.

The syntactic similarity between query and file expressions and the terminological convenience of using the terms "query expression", "query language", and "query structure" in theoretical discussions suggest that query and file languages are both structure languages in exactly the same sense. But this is obviously not the case because of the partial asymmetry between file and query language. A concrete full expression in GENSAL describes a concrete structure stored in a file,

whereas a full query formulation in the corresponding query language gives instructions to the search program on those conditions which must be satisfied by a full file structure (or by a full file expression) in order to be a "hit". From a theoretical point of view, therefore, different concepts of comprehending a query language as a linguistic phenomenon are possible. But the different concepts differ only in the theoretical and terminological respects, they make no difference in practice in using query languages.

From an extremely theoretical point of view, a query language may be seen as a special kind of metalanguage used for describing and selecting, by means of a particular query formulation in a particular search, a defined set of expressions of an object language. This set of full file expressions required by the search is potentially infinite in most searches; it contains all possible file expressions that are defined as hits by the query. Application of this query formulation to a finite database causes a finite set of file expressions to be retrieved. This concept is rather simple but does not reflect the usual concept of a search as a comparison of structures.

From another theoretical point of view, which may seem better suited to the pragmatic situation perceived by the user, a query language is a structure language. An expression in the surface query language then denotes a query extension which is a query structure. This query structure is to be compared with the file structures.

If this concept is adopted, the question inevitably arises as to whether a change in the required matching-relation (e.g., from (a) to (b) in the examples discussed above)—which implies a change of the answer set of *file* structures retrieved—also implies a change in the *query* structure itself. The question can be answered in a satisfactory and consistent manner only by making two distinctions:

> The extension denoted by a query notation must be distinguished from the intension expressed, i.e., the mere query *structure* must be distinguished from the *concept* of this query structure as a structure to be compared with file structures according to special, optionally particularized search purposes.
>
> The *query* structure, i.e., the extension denoted by the *requiring* query notation, must be distinguished from the file structure(s) *required*, i.e., from the distinct extension(s) denoted by the distinct file notations *required*. A search in a large database of generic full structures may retrieve many required file structures for one query structure.

Searches for full file structures matching with full query structures in databases containing generic (and in addition, optionally specific) full structures, may be conceived then in the following manner:

> Different matching-relations required for a particular query structure are expressed by different notations, e.g., by "$(CH_3)_3$C-OH" as a simple example of a full query expression according to (a), by "R-OH; R = *t*-butyl TO *t*-alkyl" as a simple example of a full query expression according to (b), and by the respective notations according to the examples of (c) and (d). These notations express different intensions, which give different operational information on the proper treatment of the specific query structure in comparison operations. But these different notations denote the same extension, i.e., the same specific query *structure*. And these different intensions determine the same specific query structure only in a different manner of conceiving it as a specific *query* structure which may match with specifically described (i.e., non-h-variant) matching-paths through file structures or with generically described (i.e., h-variant) matching-paths.
>
> These different intensions, determining the *same* specific *query* extension, define different sets of required file structures which may possibly exist, i.e., *different sets* of distinct *file* extensions, by giving information on the matching-relation required. In searches for file structures that may be generic structures, each of these different sets of required file structures is a potentially infinite set. [For the full query notation according to (a), for example, all conceivable distinct full file structures of the type "R-OH; R = *t*-butyl/..." or of the type "$(CH_3)_3$C-R; R = OH/..." or of yet other types, e.g., "$(CH_3)_2$C(OH)R; R = CH_3/...", are required file structures; and for the full query notation according to (b), all the distinct file structures of the type "R-OH; R = *t*-alkyl ⟨4-...⟩" or of the type "R-OH; R = *t*-alkyl ⟨4–8⟩/..." are required.] Searches in a finite database, of course, retrieve only finite subsets of these potentially infinite sets. Most of the distinct full file structures required by query notations according to (a), (b), (c), or (d) are extensionally different classes of specific structures, but they all have at least one specific full structure in common, which is identical with the query structure in the case of the specific full query structures discussed here.

In the case of generic full query expressions (e.g., "R-OH; R = *t*-alkyl! ⟨4–8⟩" according to (e), "R-OH; R = *t*-butyl/phenyl", or "$(CH_3)_3$C-R; R = OH/halogen"), denoting generic full query structures (i.e., classes of specific full structures), the file structures required may even have no specific file structure or structures in common. But each of the required file structures has at least one specific full structure in common with the generic full query structure.

The applicability of this concept to full query expressions and to full file expressions, as discussed here, implies its applicability, by analogy, to partial expressions within these full (query and file) expressions and its applicability to query expressions that are treated as expressions of substructures. It seems to be a merely terminological question whether these substructures should be conceived as partial query structures or as full query structures or even as a special kind of query structure used in a special kind of search. In the usual pragmatic view of the user (using, e.g., the Markush DARC or MARPAT systems), the same query structure may be used in a substructure search and in a search excluding the matching-relation of substructural embedment. Furthermore, the answer set of the last search would be a subset of that of the substructure search. For the sake of terminological uniformity, therefore, substructures should be conceived as full query structures, being compared with full file structures in regard to the special matching-relation of substructural embedment.

This concept, viewing query languages as structure languages, is not quite simple, but it seems to be the only theoretical way of comprehending a search algorithm with regard to a particular matching-relation in a familiar manner, viz., as a comparison of (partial) structures. The technical aspects of such comparisons are discussed in the following papers in this series.

ACKNOWLEDGMENT

We gratefully acknowledge financial support for this work from International Documentation in Chemistry—IDC GmbH, Chemical Abstracts Service, Derwent Publications Ltd., and Questel S.A. We also thank the reviewers for useful comments, which added to the clarity of presentation.

REFERENCES AND NOTES

(1) Downs, G. M.; Gillet, V. J.; Holliday, J. D.; Lynch, M. F. Computer Storage and Retrieval of Generic Chemical Structures in Patents. 10.

The Generation and Logical Bubble-Up of Ring Screens for Structurally Explicit Generics. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 215-224.
(2) Shenton, K.; Norton, P.; Langdon, M. L.; Fearns, E. A. Graphical Retrieval of Patent Information. In *Proceedings of the 9th International Online Meeting, Learned Information*, London, Dec 1985; Learned Information: Oxford, 1986.
(3) Shenton, K.; Norton, P.; Fearns, E. A. Generic Searching of Patent Information. In *Chemical Structures. The International Language of Chemistry*; Worr, W. A., Ed.; Springer-Verlag: Berlin, 1988; pp 169-178.
(4) Fisanick, W. Storage and Retrieval of Generic Chemical Structures in Patents. U.S. Patent 4642762, Feb 1987.
(5) Fisanick, W. The Chemical Abstracts Service Generic Chemical (Markush) Structure Storage and Retrieval Capability. 1. Basic Concepts. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 145-154.
(6) Stiegler, G.; Maier, B.; Lenz, H. Automatic Translation of GENSAL Representations of Markush Structures into GREMAS Fragment Codes at IDC. Presented at the Fall ACS Meeting, Washington, DC, 1990.
(7) Barnard, J. M.; Lynch, M. F.; Welford, S. M. Computer Storage and Retrieval of Generic Chemical Structures in Patents. 2. GENSAL. A Formal Language for the Description of Generic Chemical Structures. *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 151-161.
(8) Barnard, J. M.; Lynch, M. F.; Welford, S. M. Computer Storage and Retrieval of Generic Chemical Structures in Patents. 4. An Extended Connection Table Representation for Generic Structures. *J. Chem. Inf. Comput. Sci.* **1982**, *22*, 160-164.
(9) Barnard, J. M.; Lynch, M. F.; Welford, S. M. Computer Storage and Retrieval of Generic Chemical Structures in Patents. 6. An Interpreter Program for the Generic Structure Language GENSAL. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 66-70.
(10) Welford, S. M.; Ash, S.; Barnard, J. M.; Carruthers, L.; Lynch, M. F.; von Scholley, A. The Sheffield University Generic Chemical Structures Project. In *Computer Handling of Generic Chemical Structures*; Barnard, J. M., Ed.; Gower Publishing Company Ltd.: Aldershot, U.K., 1984; pp 130-158.
(11) Gordon, J. E.; Brockwell, J. C. Chemical Inference. 1. Formalization of the Language of Organic Chemistry: Generic Structural Formulas. *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 117-134.
(12) Morris, C. W. *Foundations of the Theory of Signs*, 8th ed.; University of Chicago Press: Chicago, IL, 1953.
(13) Gillet, V. J.; Downs, G. M.; Ling, A. I.; Lynch, M. F.; Venkataram, P.; Wood, J. V.; Dethlefsen, W. Computer Storage and Retrieval of Generic Chemical Structures in Patents. 8. Reduced Chemical Graphs and Their Applications in Generic Chemical Structure Retrieval. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 126-137.
(14) Welford, S. M.; Barnard, J. M.; Lynch, M. F. Computer Storage and Retrieval of Generic Chemical Structures in Patents. 5. Algorithmic Generation of Fragment Descriptors for Generic Structure Screening. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 57-66.
(15) Wittgenstein, L. Philosophische Untersuchungen. Blackwell: Oxford, 1953.
(16) Carnap, R. *Meaning and Necessity, A Study in Semantics and Modal Logic*. University of Chicago Press: Chicago, IL, 1947.
(17) Frege, G. Uber Sinn und Bedeutung. *Z. Philosophie Philosophische Kritik* **1892**, *100*, 25-50.
(18) Peirce, C. S. In *Collected Papers*; Hartshone, C., Weiss, P., Eds.; Harvard University Press: Cambridge, MA, 1931-1935; Vols. I-VI.
(19) Mockus, J.; Stobaugh, R. E. The Chemical Abstracts Service Chemical Registry System. 7. Tautomerism and Alternating Bonds. *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 18-22.
(20) Welford, S. M.; Barnard, J. M.; Lynch, M. F. Computer Storage and Retrieval of Generic Chemical Structures in Patents. 3. Chemical Grammars and their Role in the Manipulation of Generic Chemical Structures. *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 161-168.
(21) Church, A. *Introduction to Mathematical Logic*; Princeton University Press: Princeton, 1967; Vol. 1.

Computer Storage and Retrieval of Generic Chemical Structures in Patents. 12. Principles of Search Operations Involving Parameter Lists: Matching-Relations, User-Defined Match Levels, and Transition from the Reduced Graph Search to the Refined Search

WINFRIED DETHLEFSEN
BASF, Ludwigshafen/Rhein, Germany

MICHAEL F. LYNCH,* VALERIE J. GILLET, GEOFFREY M. DOWNS, and JOHN D. HOLLIDAY
Department of Information Studies, University of Sheffield, Sheffield S10 2TN, England

JOHN M. BARNARD
Barnard Chemical Information Ltd., 46 Uppergate Road, Stannington, Sheffield S6 6BX, England

Received November 16, 1990

This paper continues the establishment of a consistent framework for discussing and treating representations of generic structures described in Part 11 of this series (see preceding paper in the issue). In this part, the nature of search operations and the use of parameter lists representing both specifically and generically described parts within generic full structures as the basis for the matching operation prior to the refined search on the ECTR are considered. The nature of the matching-relations is identified, together with the concept of matching-paths for query structures in file structures. The possibility of extension of the matching operations in order to implement user-defined levels of search and a refined search are also discussed.

1. INTRODUCTION

The complexities associated with the retrieval of generic structures require a number of different levels of representation in order to provide efficient search. Consequently, several levels have been implemented, including fragment[1] and ring screens, and reduced graphs together with the use of parameters. The exact order of application of these screens is a matter of technical efficiency and may vary according to query and file characteristics. Ring screens have already been reported for the representation of full structures.[2] Reduced graphs were introduced as an additional screen,[3] but their role is now seen as being two-fold: both as screens and also as providing a preliminary mapping between the query structure and file structures, enabling more detailed comparisons to be carried out on the corresponding query and file partial structures via their ECTR representations. Parameters have so far been described for the representation of generically described partial structures, but in fact, they also provide a universally applicable concept that may be applied to other representations as well, as described below.

2. THE UNIVERSAL APPLICABILITY OF PARAMETER LISTS

As noted earlier, structural information relating to generic

0095-2338/91/1631-0253$02.50/0