

Automatic Interpretation of the Texts of Chemical Patent Abstracts. 1. Lexical Analysis and Categorization

G. G. Chowdhury and M. F. Lynch*

Department of Information Studies, University of Sheffield, Sheffield S10 2TN, England

Received February 18, 1992

A semiautomatic method for converting to GENSAL those parts of Derwent Publications Ltd. Documentation Abstracts which specify generic structures is reported in this paper and that which follows. Techniques of natural language processing (NLP) applied in a prototype system are discussed. This paper deals with the lexical isolation and categorization of tokens from the generic structure textual descriptions. Templates for processing of both the variable and multiplier expressions, which predominate, have been identified; they provide the basis for further analysis. Rules for the isolation of tokens are discussed and illustrated. Some categories of tokens are identified by morphological analysis, while others are dealt with by dictionary lookup. The output is a list of tokens along with a number of associated semantic features which help at the processing stage discussed in the following paper.

1. INTRODUCTION

Patents are an invaluable source of information on which industries depend for protection of the products of their research.¹ Patent abstracts, exemplified particularly by the *World Patent Index* published by Derwent Publication Ltd., are much more concise than patents, which are published in many different languages and often involve long and complex texts, and they provide the essential information from each patent in an easily assimilable fashion. Derwent's *World Patent Index* contains over 4 million records and grows at a rate of 12 000 patents a week.¹ Sections B, C, and E of the Central Patents Index (CPI) of Derwent cover chemical patents, and a recent estimate shows² that nearly 700 basic chemical patents are Markush DARC indexed in each weekly issue of CPI.

Chemical patents often include generic chemical structures which define families of compounds for which protection is sought, in addition to one or more individual specific substance. A generic structure generally comprises an invariant part together with associated variable groups. The variable groups represent the possible alternatives and are expressed in patents as partial structure diagrams, line notations, generic and specific radical names, and other textual expressions. The structural characteristics of generic structures, or Markush structures as they are also known, have been widely discussed, and much research and development has been work published.³⁻⁹

Until recently, generic structure databases were searchable only on the basis of fragmentation codes,¹⁰ e.g., Derwent's Chemical Code¹¹ and the IDC GREMAS code.¹² Two searchable databases of generic structures, **Markush-DARC**, developed jointly by Derwent Publications Ltd., Questel S.A., and INPI (the French Patent Office) and **MARPAT**, from Chemical Abstracts Service, have recently become available for use,⁵ while International Documentation in Chemistry (IDC) is using GENSAL for database creation and GREMAS code generation purposes.¹³ Efficient methods for creation of these databases have thus become important.

The tasks involved in this process may be divided into two phases. In the first, an analyst knowledgeable both in chemistry and in patents analyzes and interprets each description in order to arrive at the definition of the generic structure. In the second phase, the information is translated

into a representation of some kind for storage and processing. Those sections of *World Patent Index* which deal with generic structures provide a form of expression that is relatively concise when compared with the language used in patent texts. Thus, in Sections B, C, and E of Derwent's *Documentation Abstracts*, the generic structure diagram is followed by a number of statements, which we term **assignment statements**. The variable parts of the generic structure are declared by means of these assignment statements. Figure 1 shows a typical patent abstract, while Figure 2 shows some assignment statements and their corresponding GENSAL expressions. These GENSAL expressions are produced by an analyst during the process of database creation.

The work described here was aimed at determining the feasibility of automatic (or semiautomatic) translation of those parts of Documentation Abstracts which describe generic structures, with GENSAL as the target representation; a preliminary report has already appeared.¹⁴ This paper, presented in two parts, describes the capabilities of a prototype system in greater detail; the first part describes the isolation and categorization of textual tokens in patent abstracts, while the second discusses processing techniques and assesses the results of the study.

2. NATURAL LANGUAGE PROCESSING (NLP) AND SUBLANGUAGE ANALYSIS

Recent trends in NLP as applied to information retrieval have been discussed by Smeaton,¹⁵ Jacobs and Rau,¹⁶ Warner,¹⁷ Doszkocs,^{18,19} and Grishman.²⁰ Those which relate to patents include knowledge extraction, access to patent databases through significant keywords,²¹ and automatic indexing of patent databases.²² Characteristically in many NLP studies, a particular specialized domain is selected in order to reduce the complexity of the operations, often termed sublanguage analysis,^{23,24} as illustrated below with particular reference to areas of chemical information.

Ledwith²⁵ has described work on the development of a concept-oriented database for online retrieval, its purpose being to create a database with explicit conceptual information for each document in the database. The data selected was a volume of Chemical Abstracts index containing 238 000 document citations. Zamora and Blower^{26,27} studied the extraction of reaction information from the descriptions of

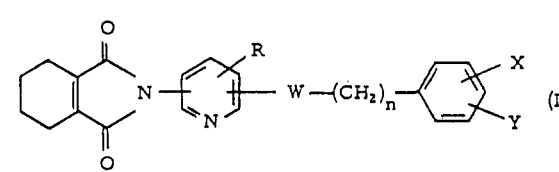
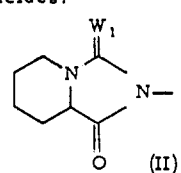
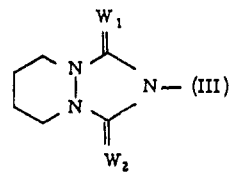
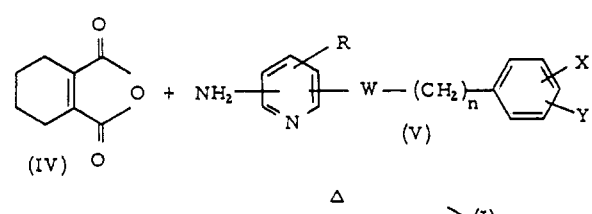
<p>83-788876/41 C02 ZOECON CORP 12.02.82-US-348555 (+US-273834) (27.09.83) A01n-43/40 C07d-401/04 N-Phenyl-alkoxy-pyridinyl -tetrahydro phthalimide derivs. - useful as pre- and post-emergent herbicides</p> <p>C83-099595</p> <p>N-((Phenylalkyl-oxy or -thio)pyridyl-3,4,5,6-tetrahydrophthalimide derivs. of formula (I) are new.</p>  <p>(I)</p> <p>(n is 1, 2 or 3; R is H, F, Cl, Br, or I; W is O or S; and X and Y are H, F, Cl, Br, I, or alkyl, alkoxy, thioalkyl, haloalkyl or haloalkoxy each having 1-6C).</p> <p>USES</p>	<p>C(6-D3, 6-D5, 6-D8, 12-P5) u</p> <p>(I) are pre- and post-emergent herbicides.</p> <p>WIDER DISCLOSURE (I) in which the tetrahydrophthalimide gp. is replaced by a gp. of formula (II) or (III) are also disclosed as herbicides.</p>  <p>(II)</p>  <p>(III)</p> <p>(W₁ and W₂ are O or S).</p> <p>PREPARATION</p> <p>US4406690-A+</p>
 <p>(IV) + (V) $\xrightarrow{\Delta}$ (I)</p> <p>(I; W = O) may also be obtd. by reaction of a hydroxy-nitropyridine with a phenylalkyl halide, followed by redn. of the NO₂ gp.</p> <p>EXAMPLE A mixt. of 0.4 g (IV), 0.6 g 5-amino-2-(p-chlorophenylethoxy)pyridine and 20 ml acetic acid was refluxed for 1.5 hrs., cooled and concd. to one-half vol. The solid was sepd. and washed with 10% EtOAc/hexane to give N-(2-(4-chlorophenylethoxy)-5-pyridyl)-3,4,5,6-tetrahydrophthalimide. (6pp1248GHDwgNo0/0).</p> <p>US4406690-A</p>	

Figure 1. Sample documentation abstract.

Assignment statements	Gensal representations
R is 1-8C alkyl, 3-8C cycloalkyl, 2-4C alkenyl, 1-4C haloalkyl, benzyl, alkoxy-methyl, or Ph (opt. subst. by halogen, methoxy, NO ₂ or carbalkoxy);	R = alkyl<1-8>/cycloalkyl<3-8>/alkenyl<2-4>/alkyl<1-4> SB halo/benzyl/methyl SB alkoxy/ Ph OSB (halogen/methoxy/NO ₂ /carbalkoxy);
R1 is H, halogen, CF ₃ , alkoxy, carbalkoxy or CN;	R1 = H/halogen/CF ₃ /alkoxy/carbalkoxy/CN;
R2 and R3 are each H, halogen, 1-4C alkyl, CF ₃ , alkoxy, carbalkoxy or CN;	R2,3 = H/halogen/alkyl<1-4>/CF ₃ /alkoxy/carbalkoxy/CN;
R4 and R5 are each 1-4C alkyl;	R4,5 = alkyl<1-4>;
X is O or S.	X = O/S;

Figure 2. Assignment statements and their representation in GEN-SAL.

syntheses of organic substances. The system was tested on 40 synthetic paragraphs from the experimental sections of papers in *The Journal of Organic Chemistry*, 36 of which were processed satisfactorily.²⁷ Using Zamora's work as a model, Ai et al.²⁸ developed a system that generates a summary

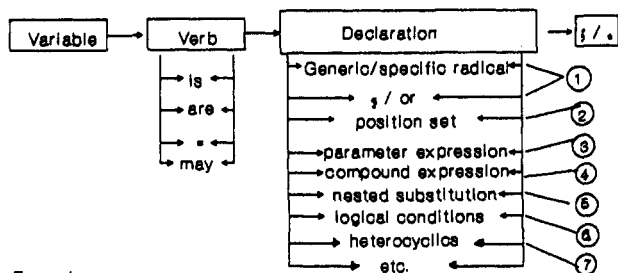
of preparative reactions from the experimental section of *The Journal of Organic Chemistry* papers. For simple synthesis paragraphs, the program produces successful results in 80-90% of the cases, but the success rate falls to 60-70% in the case of complex paragraphs.²⁸

3. SAMPLE DATABASE

The sample database used in our work consists of 73 Documentation Abstracts from Sections B and C of Derwent's *Basic Abstracts Journal* of 1984. These abstracts contain a total of 545 assignment statements, which, in combination with structure diagrams, define the generic structures. Those parts of the abstracts that comprise assignment statements were extracted to create files. Each of these files was then processed, and the corresponding output was stored in separate files. Each output file was compared with the corresponding input file to assess how far the input assignment statements were processed correctly in each case, i.e., how closely the output corresponded with the GEN-SAL statements, which a skilled analyst would have produced.

4. MODELS OF ASSIGNMENT STATEMENTS

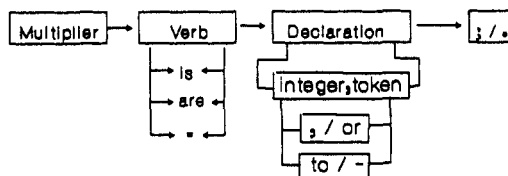
The approach is based on identification of common patterns of occurrence of assignment statements in the abstracts. 200



Examples :

1. R1 is H, Me, Et, propyl or alkyl;
2. R1 is 2-, 3- or 4-pyridyl opt. substd. by 1 or 2 lower alkyl gps.;
3. R2 is 1-6C aliphatic hydrocarbonyl;
4. R1 and R3 are halogen or 2-6C alkoxy-carbonyl;
5. R7 is 1-3C alkyl opt. substd. by OH, SH or OCH3;
6. R2 is 8C alkyl when both R1 and R3 are H;
7. R3 is lower alkyl or a 5, 6 or 7-membered heterocyclic ring;

Figure 3. Templates for variable expressions.



Examples :

- n is 0, 1 or 2;
 m and n are 0, 1 or 2;
 m = 2-5;
 n is an integer 1-3;
 p is an integer 1 to 3;

Figure 4. Templates for multiplier expressions.

Documentation Abstracts from Sections B and C of the Derwent's *Basic Abstracts Journal* were studied to identify the basic expression patterns. The most common pattern was found to be a sequence of assignment statements describing either the value of R groups, however designated, or multiplier variables. While some of the expressions were quite simple, some were embedded in more complex constructions, involving partial structure diagrams, conditional expressions, anaphoric references, etc. These observations gave rise to templates, which formed the basis of further NLP studies. The prototype was based on an expectation-driven approach, involving complex routines because of the complex nature of the data source. Figures 3 and 4 illustrate the types of template for variable and multiplier expressions.

5. PROCESSING STAGES

The prototype takes each assignment statement and produces the corresponding GENSAI output. However, various intermediate processes are involved. The input is passed to a lexical analysis phase, followed by a dictionary lookup procedure, resulting in a list of isolated and categorized tokens that are then processed syntactically and semantically. This paper describes the lexical analysis and categorization parts of the work, the analysis and interpretation parts being described in the following paper.

6. LEXICAL ANALYSIS

In general, the lexical analysis phase serves to isolate each token; this process is often coupled with the use of a lexicon to validate the tokens. It may also categorize the tokens with

a view to identifying the role of each token. The significance of the lexical analysis phase in the analysis of chemical patent abstracts is described in the following sections with examples from the sample database.

7. LEXICAL ISOLATION

The language used in the assignment statements includes chemical names, a restricted range of natural language expressions, and punctuations. The chemical names found here include descriptions of both specific and generic radicals and their combinations, for example, 'chlorophenyl', 'alkoxyaminocarbonyl', and 'phenyl(1-3C)alkyl', rather than full substance names. In addition, complex expressions such as 'NR1R2' can be found. No attempt at comprehensive analysis of chemical radical names is made here; this is a task for nomenclature translation.^{29,30} However, some analysis of compound radical names has been made.

The lexical analysis phase performs three major functions. It isolates the tokens of each assignment statement by using token-terminating conditions. Secondly, it acts as a preprocessor to standardize some tokens, especially parts of compound token expressions. Thirdly, it automatically identifies characteristic features of certain tokens and thus assigns categories to them, thereby simplifying the dictionary lookup phase. Identification of the basic categories of tokens is guided by the templates.

7.1. Token Terminators. Token terminators were identified to isolate the tokens. The isolation routine takes each assignment statement and reads the characters until one of the token-terminating conditions appears. Altogether 29 token-terminating conditions were formulated on the basis of observations of around 200 Documentation Abstracts. The broad categories of such conditions are detailed below.

- Common delimiters, 5 conditions
- Compound expressions, 4 conditions
- Position set expressions, 1 condition
- Variable expressions, 5 conditions
- Prefix matching, 5 conditions
- Suffix matching, 9 conditions

The common delimiters include space, period, comma, semicolon, etc. However, owing to the complexities introduced by chemical names, the rules were formulated in such a way that they did not fragment chemical names. Thus, in the expression 'R is methyl, ethyl, propyl ...', the system isolates each token and punctuation producing 'R', 'is', 'methyl', ',', 'ethyl', etc., but it does not isolate the constituent parts of chemical radical names like '2,2-bis(hydroxycarbonyl)ethyl'.

Compound token expressions may include parameter expressions, the component parts of which need to be isolated. Three sets of rules were formulated for this purpose, one based on prefix matching, a second on suffix matching, and the third intended for tokens involving parameter expressions. Examples of the last kind of compound expressions are phenyl-(1-5C)alkyl, phenyl(1-3C)alkyl, aryl(1-25C)alkyl, etc. The terminating conditions for isolating compound expressions split the component tokens at appropriate places. Prefix and suffix matching rules match prefixes like bi, di, tri, poly, halo, etc. and suffixes like oxy, yl, ino, ene, etc. to isolate compound token expressions like aminocarbonyl, monochlorophenyl, biphenylalkoxy, alkylamino, alkoxyalkyl, trihaloalkyl, etc.

Expressions like 'R1 is 2-, 3- or 4-pyridyl ...' denote an example of position sets. The terminating condition for dealing with position sets isolates the tokens in this example producing '2', ',', '3', 'or', '4', 'pyridyl', etc. [the hyphens are dropped by a standardization procedure which acts as a preprocessor (see

Section 7.2) and strips off some part(s) of tokens which may not be necessary for further processing]. The rule can also distinguish between a position set and a chemical name, for example, it does not isolate the constituent parts of the chemical name '1,3-dioxan-2-yl'.

Variable expressions, which occur as **R1**, **R5**, **R23**, **Ra**, **Rb**, **R'**, **X1**, **R1-R5**, etc., are readily identified. The terminating conditions enable the integer (or equivalent) part of the variable expression to be isolated from the alphabetic part.

7.2. Preprocessing of Tokens. Thirteen rules were formulated for preprocessing the tokens. Most of these rules identify the occurrence of a particular type of expression and strip off one or more constituent characters while writing it to the output file and were designed for ease of handling the tokens at the processing stage. These rules isolate 'C' from tokens like **2-3C alkyl**, or ')' from tokens like **(alkyl)amino**, etc. The preprocessor also identifies occurrence of tokens like **alk(en)yl** and produces two tokens **alkyl** and **alkenyl**.

7.3. Categorization of Common Tokens. Rules to determine the occurrence of the most commonly occurring categories of tokens include the following:

- Variable names like 'R1', 'Ra', 'X1', etc.
- Multiplier names like 'm', 'n', 'p', etc.
- Parameter expressions like '2-3C' in '2-3C alkyl'
- Integer expressions like '2-5' in 'm is an integer 2-5'
- Qualifier expressions like 'lower' in '... a lower alkyl, alkenyl ...'
- Verbs like 'are' in 'R1 and R2 are ...'
- Complex expressions like 'NR7R8' in 'R3 can be NR7R8'
- Line notations like '2,6-Me2-4-propargyl' in 'R5 is ... 2,6-Me2-4-propargyl, 2,6-Me2-4-allyl or CH2Z-CH=CCl-CH2C6H4T'

Along with the tasks of token isolation and preprocessing, the lexical isolation phase also looks for these common categories of tokens; when identified, the corresponding category is written along with the token. This reduces the load of the dictionary lookup procedure where each token is assigned the appropriate token category. The output of the lexical isolation phase is thus a list of isolated tokens with some of the tokens categorized. This forms the input to the token categorization or dictionary lookup phase.

8. DICTIONARY LOOKUP PHASE

In brief, the dictionary lookup phase performs two major functions. It attaches one or more characteristic features to the given input tokens. These features provide guidelines for decisions at the processing phase where the tokens are processed into GENSAL statements. For attaching the appropriate characteristic feature to a given token it is passed to two dictionaries—a dictionary of chemical tokens and a dictionary of nonchemical tokens. Secondly, this phase allows the user to incorporate any token interactively into the dictionary.

The data structure of the dictionary of chemical tokens is simple and is intended to handle most general cases rather than to be comprehensive. It focuses particularly on features of nomenclatural terms, differentiating between generic and specific terms, and identifying characteristic features which help in processing the tokens. Each item in the dictionary of chemical tokens is a record with a maximum of seven fields. The first field contains the token itself, while the other fields denote specific features which characterize the given token. For example, one feature denotes the type of token, i.e., whether the token is a generic or a specific radical, a variable expression or a complex notation, etc. If the token is a generic radical

then the subsequent features denote whether it is a cyclic or an acyclic radical; if cyclic, then whether it is aromatic, heterocyclic, or alicyclic; whether the radical can have a variable range of carbon atoms, and if so, what is the minimum number of carbon atoms, etc. The tokens are thus characterized by a hierarchy of features, each representing some semantic aspect of the token and therefore useful for semantic processing. The output of the dictionary lookup phase is thus a list of all the tokens occurring in an assignment statement, with associated characteristic features.

8.1. Dictionary Updating. If a token is not found in the dictionary of chemical tokens, it is passed to the nonchemical token dictionary. If, however, a token does not occur in either of the dictionaries, the system enables immediate updating to be performed.

9. OUTPUT

The two phases of the prototype, token isolation and categorization, take the assignment statements from the input Documentation Abstracts and produce a list of tokens associated with up to six characteristic features. The internal processes include isolation and preprocessing of tokens followed by categorization in two phases. The user can interact with the system to add new tokens to the dictionary. Syntactic and semantic processing then process the list of categorized tokens, i.e., the output of the lexical categorization phase. These processing stages are discussed in the following part with an evaluation of the results.

ACKNOWLEDGMENT

The funding of this work by the Commonwealth Scholarship Commission, U.K., is gratefully acknowledged. We thank Derwent Publication Ltd. for providing materials for this work.

REFERENCES AND NOTES

- (1) Evers, H. Patent information. In *Perspectives in Information Management 1*; Oppenheim, C. L., Ed.; Butterworths: London, 1989; pp 219–256.
- (2) *Derwent Online News*, No. 3, Sept 1991, 3.
- (3) Simmons, E. S. The grammar of Markush structure searching: vocabulary vs syntax. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 45–53.
- (4) Sibley, J. F. Too broad generic disclosures: a problem for all. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 5–9.
- (5) Wilke, R. N. Searching for simple generic structures. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 36–40.
- (6) 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.
- (7) Milne, G. W. A. Very broad Markush claims; a solution or a problem: Proceedings of a round-table discussion held on August 29, 1990. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 9–30.
- (8) Dethlefsen, W.; Lynch, M. F.; Gillet, V. J.; Downs, G. M.; Holliday, J. D. Computer storage and retrieval of generic chemical structures in patents. 11. Theoretical aspects of the use of structure languages in a retrieval system. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 260–270.
- (9) Lynch, M. F.; Downs, G. M.; Gillet, V. J.; Holliday, J. D.; Dethlefsen, W. Generic chemical structures in patents—an evaluation of the Sheffield University research work. In *Proceedings of the Montreux 1989 International Chemical Information Conference*; Collier, H. R., Ed.; Springer-Verlag: Berlin, 1989; pp 161–173.
- (10) Cloutier, K. A. A comparison of three online Markush databases. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 40.
- (11) Shenton, K.; Norton, P.; Ferns, E. A. Generic searching of patent information. In *Chemical structures: the international language of chemistry*; Warr, W., Ed.; Springer: Berlin, 1988; pp 169–178.
- (12) Stiegler, G.; Maier, B.; Lenz, H. Automatic translation of GENSAL representations of Markush structures into GREMAS fragment codes. Presented at the Second International Conference on Chemical Structure Languages, Noordwijkerhout, The Netherlands, 1990.
- (13) Bernard, J. M. A comparison of different approaches to Markush structure handling. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 64–68.
- (14) Chowdhury, G.; Lynch, M. F. Natural language processing of the texts of chemical patent abstracts. In *Intelligent Text and Image Handling*,

- Proceedings of RIAO'91 Conference*, Barcelona, Apr 2-5, 1991; Lichnerowicz, A., Ed.; pp 740-753.
- (15) Smeaton, A. F. Information retrieval and natural language processing. In *Prospects for Intelligent Retrieval: Proceedings of Informatics 10*, Jones, K. P., Ed.; Aslib: London, 1990; pp 1-14.
 - (16) Jacobs, P. S.; Rau, L. F. Natural language techniques for intelligent information retrieval. In *Eleventh International Conference on Research and Development in Information Retrieval*, Grenoble, France, 1988; ACM: Washington, 1988; pp 85-99.
 - (17) Warner, A. J. Natural language processing. In *Annual Review of Information Science and Technology*; Williams, M., Ed.; Elsevier Science Publishers: London, 1987; Vol. 22, pp 79-108.
 - (18) Doszkocs, T. E. Natural language processing in information retrieval. *J. Am. Soc. Inf. Sci.* **1986**, *37*, 191-196.
 - (19) Doszkocs, T. E. IR, NLP, AI and UFOS: or IR, relevance, natural language problems, artful intelligence and user-friendly online system. In *Proceedings of the 9th International Conference on Research and Development in Information Retrieval*, Pisa, Italy, 1986; Rabitti, F., Ed.; ACM: Washington, 1986; pp 49-53.
 - (20) Grishman, R. Natural language processing. *J. Am. Soc. Inf. Sci.* **1984**, *35*, 291-296.
 - (21) Turner, W. A.; Buffet, P.; Laville, F. LEXITRAN for an easier public access to patent database. In *Intelligent Text and Image Handling, Proceedings of RIAO'91 Conference*, Lichnerowicz, A., Ed.; Barcelona, Apr 2-5, 1991; pp 320-336.
 - (22) O'Hara, M. P.; Pagis, C. The PHARMSEARCH Database. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 59-63.
 - (23) Liddy, E.; Jorgenson, C.; Sibert, E.; Yu, E. S. Processing natural language for an expert system using a sublanguage approach. In *Intelligent Image and Text Handling, Proceedings of the RIAO'91 Conference*, Lichnerowicz, A., Ed.; Barcelona, Apr 2-5, 1991; pp 707-717.
 - (24) Kitteredge, R.; Lehrberger, J. *Sublanguages: Studies of Languages in Restricted Semantic Domains*; Walter DeGruyter: Berlin, 1982.
 - (25) Ledwith, R. Development of a large concept-oriented database for information retrieval. In *Eleventh International Conference on Research and Development in Information Retrieval*, Grenoble, France, 1988; ACM: Washington, 1988; pp 651-661.
 - (26) Zamora, E.; Blower, P. E. Extraction of chemical reaction information from primary journal text using computational linguistics techniques. 1. Lexical and syntactic phases. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 176-181.
 - (27) Zamora, E.; Blower, P. E. Extraction of chemical reaction information from primary journal text using computational linguistics techniques. 2. Semantic phase. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 181-188.
 - (28) Ai, C. S.; Blower, P. E.; Ledwith, R. H. Extraction of chemical reaction information from primary journal text. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 163-169.
 - (29) Cooke-Fox, D. I.; Kirby, G. H.; Rayner, J. D. Computer translation of IUPAC systematic organic chemical nomenclature. 1. Introduction and background to a grammar. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 101-105.
 - (30) Cooke-Fox, D. I.; Kirby, G. H.; Lord, M. R.; Rayner, J. D. Computer translation of IUPAC systematic organic chemical nomenclature. 5. Steroid nomenclature. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 122-127.