II modems (at 300 baud). The data entry software is written in SNOBOL4, a text-oriented symbolic language.

In-house use of the partially completed database has clearly demonstrated its usefulness in environmental assessment work and toxicology research. Future incorporation of PHYTOTOX into SPHERE[5] will greatly amplify the usefulness of the database since it will then be linked to other related databases such as SANSS[6] and CHEMFATE.[12] The interactive use of these databases will be valuable in predicting how a chemical released into the environment will behave and whether or not it will come in contact with a plant species in sufficient concentration to cause an adverse or toxic effect. Realizing that such evaluations are only predictions and that testing must often follow, it is also worth noting that the PHYTOTOX database can serve a second major role in designing tailored experiments for specific chemicals. In this manner, the database is expected to prove useful in the environmental assessment of new commercial chemicals.

ACKNOWLEDGMENT

Development of this database has been supported through cooperative agreements CR-807391 and CR-810195 with the U.S. Environmental Protection Agency. Thanks are extended to the many students who have been involved with the evaluation of the research papers used in preparing this database and to Tom Weaver for his assistance with the GIPSY processing system. This article has not been subjected to the U.S. Environmental Protection Agency's peer and policy review and, therefore, does not necessarily reflect the views of the Agency, and no official endorsement of trade names or commercial products should be inferred.

REFERENCES AND NOTES

(1) Eichers, T. R. In "CRC Handbook of Pest Management in Agriculture"; Pimentel, D., Ed.; CRC Press: Boca Raton, FL, 1980; Vol. II, p 3.
(2) Pimentel, D. In "CRC Handbook of Pest Management in Agriculture"; Pimentel, D., Ed.; CRC Press: Boca Raton, FL, 1980; Vol. I, p 3.
(3) Pimentel, D.; Krummel, J.; Gallahan, D.; Hugh, J.; Merrill, A.; Schreiner, I.; Vittum, P.; Koziol, F.; Back, E.; Yen, D.; Fiance, S. In "CRC Handbook of Pest Management in Agriculture"; Pimentel, D., Ed.; CRC Press: Boca Raton, FL, 1980; Vol. II, pp 45, 46.
(4) Milne, G. W. A.; Heller, S. R. "NIH/EPA Chemical Information System". *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 204-211.
(5) "Scientific Parameters in Health and the Environment, Retrieval and Estimation: A Requirement Analysis and Examination of Alternatives"; CRC Systems Incorporated: Fairfax, VA, 1981; EPA Contract 68-01-4795.
(6) Milne, G. W. A.; Heller, S. R.; Fein, A. E.; Frees, E. F.; Margaret, R. G.; McGill, J. A.; Miller, J. A.; Spiers, D. S. "The NIH-EPA Structure and Nomenclature Search System". *J. Chem. Inf. Comput. Sci.* **1978**, *18*, 181-186.
(7) Brown, H. D.; Costlow, M.; Cutler, F. A., Jr.; Demott, A. N.; Gall, W. B.; Jacobus, D. P.; Miller, C. J. "The Computer-Based Chemical Structure Information System of Merck Sharp and Dohme Research Laboratories". *J. Chem. Inf. Comput. Sci.* **1975**, *16*, 5-10.
(8) Bond, V. B.; Bowman, C. M.; Davison, L. C.; Roush, P. F.; Young, L. F. "Applications of the Wiswesser Line Notation at the Dow Chemical Company". *J. Chem. Inf. Comput. Sci.* **1982**, *22*, 103-105.
(9) "General Information Processing System (GIPSY) Manual"; University of Oklahoma Office of Information Systems Programs: Norman, OK, 1982.
(10) Calkins, J. A.; Keefer, E. K.; Ofsharick, R. A.; Mason, G. I.; Tracy, P.; Atkins, M. "Description of CRIB, the GIPSY Retrieval Mechanism, and the Interface to the General Electric MARK III Service". *Geol. Surv. Circ. (U.S.)* **1978**, No. 755-A.
(11) Ross, R. H.; Kemp, H. T.; Pyon, M. G.; Hammons, A. S.; Ensminger, J. T. "Chemicals Tested for Phytotoxicity"; Oak Ridge National Laboratory: Oak Ridge, TN, 1979; EIS-155/V1.
(12) Howard, P. H.; Sage, G. W.; Lamacchia, A. "The Development of an Environmental Fate Data Base". *J. Chem. Inf. Comput. Sci.* **1982**, *22*, 38-44.

Central Patents Index Chemical Code: A User's Viewpoint†

EDLYN S. SIMMONS

Merrell Dow Pharmaceuticals Inc., Cincinnati, Ohio 45215

Received August 8, 1983

Subscribers to Derwent Publications Ltd.'s Central Patents Index can use the Chemical Fragmentation Code to identify patents describing a chemical compound whether its structure is expressed generically or specifically and whether the compound is claimed or merely described as novel. Compounds are encoded via molecular fragments—elements, carbon chains, rings, and functional groups—which are overcoded on a single record to define a Markush group and are retrieved by encoding chemical structures with the same fragment codes, expressing Markush groups by the Boolean OR and combining fragments with the ORBIT LINK operator. Examples are presented in which retrieval with the Chemical Fragmentation Code is compared with the retrieval of patent references by alternative search techniques.

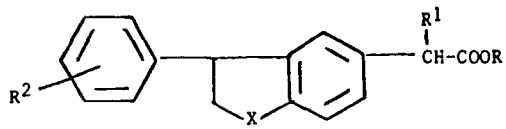
Because much of the information published about chemical compounds appears only in chemical patents, patents are valuable resources whenever chemical information is needed. Chemical patents are especially important for searchers employed by organizations involved in the development or marketing of chemical compounds and processes. Only patents can tell us whether a compound or process is protected by patent claims. Patents are also vitally important for evaluating the patentability of new inventions, for valid patents can be obtained only for inventions that have not previously been taught or suggested in the prior art, and the prior art includes all of the world's patents. But patents are useful as information sources only to the extent that they are indexed so that the compounds they describe can be identified.

Modern chemical inventions, especially those involving new chemical compounds, are typically defined in terms of complex generic structures in what is called Markush format, so indexing the compounds in patents is not as simple as indexing compounds in the journal literature. A Markush formula such as formula I in Figure 1 includes one or more variable substituents defined in terms of a group of alternative molecular substructures, which is called a Markush group. The position of substitution in a Markush formula can also be variable. The name Markush is derived from a ruling in 1925 on a patent application filed by an inventor named Eugene A. Markush,[1] in which the U.S. Patent Office declared that this format was a legally acceptable way to define a genus of compounds. A

† Presented at the 185th National Meeting of the American Chemical Society, Seattle, WA, March 24, 1983.

0095-2338/84/1624-0010$01.50/0

GENERIC FORMULA I



wherein R = H or C_{1-4} alkyl
R^1 = OH or NH_2
R^2 = H, OH, C_{1-4}, alkoxy or halogen
X = O or S

Figure 1.

Markush formula is merely a convenient shorthand method for expressing the chemical structures of a large number of specific compounds; it has no significance of its own. The prior art for a Markush formula includes every reference that overlaps with the genus to the extent that one or more of the specific compounds embodied by the genus is described by the reference. Figure 2 shows the first page of a patent that discloses a Markush formula that overlaps with formula I. The amount of overlap between the structures is better illustrated in Figure 3, where the nonoverlapping scope of the two generics has been blanked out.

We often perform a search to determine what aspects of a chemical invention are new and patentable and whether compounds that define the invention are known and perhaps patented by someone else. For these searches, every reference to the compounds is relevant to us, whether the reference is a literature citation or a patent, whether the compounds are described specifically or generically, whether the compounds are claimed in a patent or merely disclosed, and whether or not the reference teaches the intended use of the compounds.

A patentability search or a search to determine the full patent status of a compound or a genus of compounds therefore needs to be more thorough than most other types of searches. Ideally, such a search would be exhaustive without exhausting the searcher. No single source of information can guarantee access to every published reference to a genus of compounds. We use the Central Patents Index (CPI) Chemical Fragmentation Code to search for patent references to compounds to the Derwent Publications Ltd.'s Central Patents Index because no other retrieval tool allows us to search so many patents so rapidly and in so much depth.

Most indexing systems define compounds too narrowly or too broadly to allow reliable and efficient identification of references to compounds that are disclosed only generically. The Chemical Abstracts Service registry files can be searched generically for references to Markush structures, but only specific exemplified compounds can be retrieved, because only those compounds are indexed. A search for formula I by means of CAS Online does not retrieve the *Chemical Abstracts* entry that corresponds to the patent in Figure 2, because there are no specific examples in the patent of the overlapping compounds shown in Figure 3.

Unlike ordinary indexing and classification systems, the Central Patents Index Chemical Fragmentation Code was

United States Patent [19] [11] **4,341,792**
Barnish et al. [45] **Jul. 27, 1982**

[54] **HETEROBICYCLIC KETO- AND AMINO-ACIDS, ESTERS AND AMIDES**

[75] Inventors: **Ian T. Barnish, Ramsgate; Peter E. Cross, Canterbury, both of England**

[73] Assignee: **Pfizer Inc., New York, N.Y.**

[21] Appl. No.: **280,862**

[22] Filed: **Jul. 6, 1981**

Related U.S. Application Data

[63] Continuation-in-part of Ser. No. 187,431, Sep. 15, 1980, abandoned.

[30] **Foreign Application Priority Data**

Sep. 15, 1979 [GB] United Kingdom 7932049

[51] Int. Cl.³ A61K 31/34; A61K 31/35; C07D 307/84; C07D 311/04

[52] U.S. Cl. 424/275; 424/283; 424/285; 549/23; 549/57; 549/58; 549/398; 549/405; 549/406; 549/462; 549/471

[58] Field of Search 260/346.22, 346.73, 260/345.2; 424/275, 283, 285; 549/23, 57, 58

[56] **References Cited**

U.S. PATENT DOCUMENTS

4,029,811 6/1977 Tamagnone et al. 424/285
4,138,397 2/1979 Bohme 260/239.1
4,148,920 4/1979 Barnish et al. 424/319

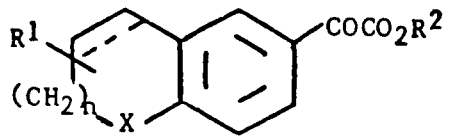
OTHER PUBLICATIONS

Chatelus, Ann. Chim., 4, 505-547, (1949); Chem. Abstr., 44, 1975c, (1950).
European Patent Appln. 8, 752 Publ. Mar. 19, 1980; Derwent Abstracts, 20590c/12 (1980).
Chatelus et al., Comptes Rendus, 224, 1777-1779, (1947).

Primary Examiner—Richard Raymond
Attorney, Agent, or Firm—Connolly and Hutz

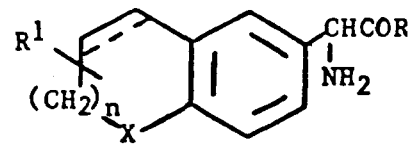
[57] **ABSTRACT**

Heterobicyclic glyoxylic acids, L- and DL-heterobicyclic glycines and their derivatives of the formulae



I

and



II

and pharmaceutically acceptable cationic and acid addition salts thereof, wherein R is OR^2 or NHR^3;

R^2 is hydrogen or alkyl having from one to four carbon atoms;

R^3 is hydrogen, alkyl having from one to four carbon atoms, alkoxyalkyl having from one to four carbon atoms in each of the alkyl groups or $R^4R^5C_6H_3CH_2$— where R^4 and R^5 are H, OH, F, Cl, Br, I, or alkyl or alkoxy having from one to four carbon atoms

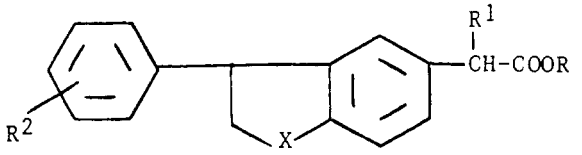
R^1 is hydrogen, alkyl having from one to four carbon atoms or $R^4R^5C_6H_3$—; X is oxygen or sulfur; n is 0 or 1 and the broken line represents an optionally present double bond; useful in treatment of diseases and conditions which are characterized by reduced blood flow, reduced oxygen availability or reduced carbohydrate metabolism in the cardiovascular system.

39 Claims, No Drawings

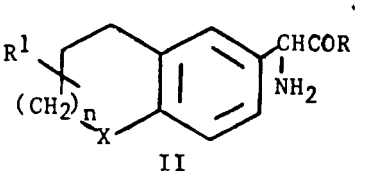
Figure 2.

GENERIC FORMULA I

U.S. 4,341,792



wherein R = H or C_{1-4} alkyl
R^1 = NH_2
R^2 = H, OH, C_{1-4}, alkoxy or halogen
X = O or S



II

and pharmaceutically acceptable cationic and acid addition salts thereof, wherein R is OR^2

R^2 is hydrogen or alkyl having from one to four carbon atoms;

R^4 and R^5 are H, OH, F, Cl, Br, I, or alkoxy having from one to four carbon atoms

R^1 is $R^4R^5C_6H_3$—; X is oxygen or sulfur; n is 0

useful in treatment of diseases and conditions which are characterized by reduced blood flow, reduced oxygen availability or reduced carbohydrate metabolism in the cardiovascular

Figure 3.

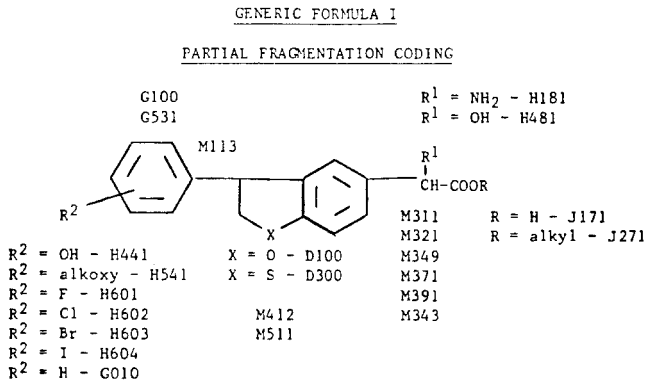


Figure 4.

expressly designed to accommodate Markush formulas. The Code reproduces chemical structures as a set of molecular fragments of the same type that patent agents and attorneys use to construct Markush formulas when we write patent applications. There are codes for functional groups, rings, carbon chains, and elements, as well as codes for formulation types, uses, and chemical reaction types.[2] Most of the code symbols that describe formula I are shown in Figure 4.

Derwent's coders encode molecular structures by listing every encodeable fragment that is permitted in the molecule. Markush groups are encoded simply by overcoding every alternative fragment on the same record so that all of the fragments are available for searching in any combination.

When a searcher wishes to retrieve references to a chemical structure, he or she encodes it in the same way. Searches may be performed online on the SDC ORBIT system in files WPI and WPIL or in-house on magnetic tapes purchased from Derwent. Access online to the fragmentation coding is restricted to CPI subscribers. The code symbols are combined online with the LINK operator to limit retrieval to a single structural record in the event that a patent has more than one encoded structure. In Figure 5 and in most of our searches, we use the symbol "+" as a synonym for LINK to simplify input. The alternative code symbols for members of Markush groups are combined by means of the OR operator before being LINKed with the remaining codes.

The strategy for retrieving references to formula I illustrated in Figure 5 retrieved 157 patent families published since 1963. These represent every patent in the CPI file that has every one of the encoded fragments, regardless of whether the search query or the patent describes a specific compound or a genus. The Fragmentation Code does not distinguish between fragments that are present together and alternative members of a Markush group and does not define every connection in an encoded molecule, so some of the references will be false drops. If we are attempting to confirm the patentability of a new compound, we hope that *all* of the references are false drops. In a patentability search, a modest number of closely related false drops is a useful thing. When the prior art fails to disclose the identical compounds one wishes to patent, the compounds may still be unpatentable if they are found to be obvious over prior art compounds. The false drops that disclose analogous compounds are extremely valuable in evaluating patentability and for confirming the accuracy of the retrieval strategy. The number of false drops one obtains from a CPI Fragmentation Code search for compounds of this type is usually quite manageable, especially when compared with the number of irrelevant patents one would retrieve by searching the traditional way, at the Patent Office Public Search Room, or by searching CPI by means of Manual Code cards.

As shown in Figure 6, a search for generic formula I at the U.S. Patent and Trademark Office Public Search Room would require evaluating the content of 2971 patents from four subclasses. Only 322 of these patent were published before the beginning of CPI's Farmdoc service in 1963, although some others will have been outside of the scope of Derwent's coverage before 1970. The CPI Manual Code also functions as a patent classification system. Searching the Manual Codes that apply to benzofurans and benzothiophenes involves scanning the abstracts of 3113 patents, 1841 of which are U.S. patents or have equivalent U.S. patents. By contrast, the Fragmentation Code search retrieved only 157 patent families, 86 of which have U.S. patents as members. By a weeding out of compounds without the substitutents required by formula

GENERIC FORMULA I

Search Logic Using CPI Fragmentation Code

SYN + FOR LINK SUBS M0,M2,M3		Postings	
	WPI	WPIL	TOTAL
SS 1: G100 + M531 + (D100 OR D300)	1518	240	1758
SS 2: 1 + (J171 OR J271) + (H481 OR H181) + (M412 OR ALL M43:/M0)	520	100	620
SS 3: 2 +NOT H2 +NOT H3 +NOT H7 +NOT H9 +NOT J3 +NOT J4 +NOT J5 +NOT J6 +NOT J9 +NOT K0	199	26	225
SS 4: 3 + M113 + M511 + M520	91	18	109
SS 5: 4 + M311 + M321 + M343 + (M370 OR M371) + M391	58	10	68
SS 6: 5 + M349	0	2	2
SS 7: 2 + M900 OR 4 + M901 OR 5 + M902 OR 6	149	8	157
SS 8: 7 + ALL P51: OR 7 + ALL P52:	21	3	24
SS 9: 7 AND PFIZER/PA	3	1	4
SS 10: 8 AND 9	0	1	1

-1-
AN - 27689D/16 (81)
TI - Bicyclic keto-and amino-acid, ester and amide cpds. - useful for treating e.g. ischaemic heart disease, cardiac failure and diabetes
DC - B02
PA - (PFIZ) PFIZER LTD
PN - EP--26593-D16 JS6053674-D26 DK8003896-D31 US4341792-E32
DS - BE CH DE FR GB IT LI LU NL SE BE CH DE FR GB IT LI LU NL SE

Figure 5.

GENERIC STRUCTURE I

U.S. Patent Classification	U.S. Patents 1789-1982	U.S. Patents 1963-1982	CPI Patents 1963-1982
424/275	1328	1225	
424/285	1435	1251	
549/58	112	86	
549/469	96	87	
Total	2971	2649	
CPI MANUAL CODES		1841	3113
CPI FRAGMENTATION CODE		86*	157*

* Basic patents through CPI week 8237

Figure 6.

I, the Code eliminated 95% or more of the patents retrieved by a traditional patent search.

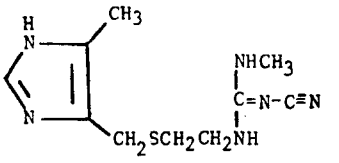
Not every patent search requires that we retrieve every published reference to a chemical structure. By use of the Fragmentation Code to search CPI, retrieval can be limited still further by adding nonstructural parameters to the search strategy. In Figure 5, statement 8 limits the previous search to patents that disclose any kind of cardiovascular activity, using all of the cardiovascular activity code symbols in truncated form. There are now only 24 patent families to review. Fragmentation Code searches in files WPI and WPIL can be combined with all of the other available retrieval parameters. By limiting this search to patents assigned to Pfizer in statement 9, we retrieve only four patents. If this search had been directed to finding patents covering compounds of formula I that have cardiovascular activity and are assigned to Pfizer, as shown in statement 10, we would have retrieved only one patent, the one illustrated in Figure 2.

The Chemical Fragmentation Code was developed in the early 1960s with the latest computer technology—punched-card sorters. The number of encodeable fragments was therefore limited to the number of punch positions available on a computer card record. The modern computer has freed the Code from the restrictions imposed by the punched card. In order to make searching with the Chemical Fragmentation Code more specific, the Code has been revised several times since 1963 to provide additional molecular descriptors. Each improvement in the Code dramatically increased the precision of the Code for searches performed in the subsequent portions of the file. Unfortunately, use of the full capability of the Code in each chronological section of the file results in a complex retrieval strategy, especially where the search covers a generic structure that is itself complex.

There are a number of simple ways to minimize the complexity of such searches. Complex generics can be subdivided and searched as several simpler structures. Even in a search for a single compound, it is never necessary to include every encodeable molecular fragment in the retrieval strategy. Application of the Code correctly is less difficult with the help of the new generation of instruction manuals, which are far more detailed than the ones they replaced, but any fragment whose coding is still uncertain can be ignored or searched by means of several alternative code symbols.

When a Markush formula is being searched, it is usually unnecessary—and sometimes impossible—to encode every possible molecular fragment. When several common functional groups are present in Markush group, it is usually best to omit the entire Markush group from the retrieval strategy. In my search for formula I, I have done exactly that; this search does not include any of the substituents on the isolated benzene ring. In most cases, omission of the coding for carbon

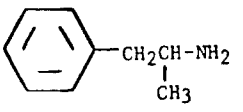
CIMETIDINE



SYN + FOR LINK SUBS M2, M3	Postings	Code Revision
SS 1: M903	30769	
SS 2: 1 + F521 + H598 + L110 + L250 + M240 + M413	27	1963
SS 3: 2 + NOT H1 +NOT H2 +NOT H3 +NOT H4 +NOT H6 +NOT H7 +NOT J0 +NOT J1 +NOT J2 +NOT J3 +NOT J4 +NOT J5 + M521 + M530 + M540	22	1970
SS 4: 3 + M210 + ALL M27: + M281 + M311 + M312 + M321 + M342 + ALL M37: + ALL M38: + M391	22	1972
SS 5: 4 + F014 + F015 + M273 + M373 + M383 +NOT H8 +NOT J9 +NOT K1 +NOT K2 +NOT K3 +NOT K4 +NOT K5 +NOT K6 +NOT K7 +NOT K8 +NOT K9 +NOT L3 +NOT L4 +NOT L5 +NOT L6 +NOT L7 +NOT L8 +NOT L9	22	1981

Figure 7.

AMPHETAMINE



SYN + FOR LINK SUBS M2,M3	Postings	Code Revision
SS 1: M903	30769	
SS 2: 1 + G100 + H181 + M414 + M531	987	1963
SS 3: 2 +NOT H2 +NOT H3 +NOT H4 +NOT H5 +NOT H6 +NOT H7 +NOT J0 +NOT J1 +NOT J2 +NOT J3 +NOT J4 +NOT J5 +NOT K0	162	1970
SS 4: 3 + M280 + M313 + M321 + M331 + M342 + M391 + ALL M37:	53	1972
SS 5: 4 + G010 + H100 + M373 +NOT H9 +NOT J9	39	1981
SS 6: CIMETIDINE/SS	22	
SS 7: 5 AND 6	1	

-1-
AN - 79611D/43 (81)
TI - Suppressing appetite in order to lose weight - by oral admin. of synergistic mixt. of amphetamine anorexiant and cimetidine
DC - B03 B05
PA - (RITT/) RITTER A
PN - US4293562-D43

Figure 8.

chain lengths and for the position of ring substitution, which are two of the trickiest parts of the Code to apply, simplifies the retrieval strategy enormously while adding few false drops to most searches and reducing the opportunity for introducing errors.

In general, the more uncommon the molecular structure being searched, the less completely it needs to be encoded for precise retrieval and the less complex the retrieval strategy needs to be. Figure 7 illustrates the molecular structure of cimetidine and the results of a search for cimetidine in the portion of file WPIL that includes the latest refinement of the Fragmentation Code. There were 30769 coded patents in that portion of the file on the day this search was run. Search statement 2 includes only the fragments that have been searchable since 1963 and results in only 27 postings. Addition of the code symbols introduced in 1970 in statement 3 reduces the number of postings to 22. Inclusion of the fragments that became searchable in 1972 and in 1981 in statements 4 and

5, respectively, adds nothing at all to the precision of this search.

On the other hand, molecules composed entirely of common chemical groups could not be defined precisely by the origial version of the Fragmentation Code. For searches involving simple aliphatic and carbocyclic molecules, the effort spent in encoding the molecule in detail is well repaid by the increase in precision the new descriptors provide. The molecule in Figure 8 is amphetamine. A search in the same portion of the file with the 1963 version of the Fragmentation Code in statement 2 yields 987 patent families. The new descriptors introduced in 1970 reduce the number of postings in statement 3 to 162. The 1972 version of the Code gives only 53 postings in statement 4, and the 1981 version gives only 39 in statement 5, which is comparable with the number of patents we retrieved for cimetidine.

This improved precision reduces the number of false drops only in the later parts of the file, of course. A search for all references to amphetamine in the whole of files WPI and WPIL would not be at all precise. Most simple molecules such as amphetamine have been known for a long time, and searches that involve such molecules usually involve some other inventive feature. If the other feature of the invention is a second compound, the search can be very precise indeed. The cimetidine search in Figure 7 was saved and recalled in statement 6, and the two searches were combined with the AND operator in statement 7. The only patent this combined strategy retrieved is directed to a composition comprising precisely the two compounds we were looking for. Had we combined the cimetidine search with statement 2, we would have obtained only one more posting, which is quite an improvement over the 987 documents retrieved by the strategy based upon amphetamine alone!

The Fragmentation Code can be used for searching the CPI files whenever molecular structure is an important feature of the invention being searched. For searches involving chemical processes, simple aliphatic and carbocyclic compounds, or nonpharmaceutical, nonagricultural compositions containing well-known compounds, it is usually not the most efficient way to search. But for complex molecules, especially heterocyclic and organometallic compounds, nothing else retrieves so many relevant references and so few false drops and does it so quickly. If you have not been using it, you probably do not know how many references you have been missing.

REFERENCES AND NOTES

(1) Ex parte Markush, 340 *OG* 839, 1925 *CD* 125.
(2) Norton, P. "Central Patents Index (CPI) as a Source of Information for the Pharmaceutical Chemist". *Drug Inf. J.* **1982**, 208-215.

Semiautomatic Indexing of Structured Information of Text

FUJIO NISHIDA,* SHINOBU TAKAMATSU, and YONEHARU FUJITA
Department of Electrical Engineering, Faculty of Engineering, University of Osaka Prefecture, Mozu-Umemachi, Sakai, Osaka, Japan 591

Received June 3, 1983

This paper presents a method of semiautomatic information extraction from text such as patent claim sentences or summaries of technical papers written in English as well as in Japanese. The input sentences are parsed, reduced, and normalized into almost the same form of the internal expressions for both the languages except terms. Subsequently, specified information is extracted in a specified language of English or Japanese.

INTRODUCTION

Automatic text processing has been studied actively with the advance of computer technology. The first main linguistic work was the automatic indexing of text, and various approaches of text analysis were established on the basis of statistical distribution of characteristic and technical terms involved in text.[1,2]

Subsequently, structural analysis of text has been introduced.[5-11] It is based on the concept of case and frame presented by Fillmore and Minsky[3,4] and aimes at processing of structured information or knowledge involved in text. Extacting, storing, and handling of structured information will be essential to knowledge engineering and information science in the near future.

This paper describes the outline of a method of semiautomatic information extraction from texts such as patent claim sentences and summaries of technical papers written in English as well as in Japanese. The information to be extracted is designated by a specification table as shown under Specification for Extracted Information. From the practical viewpoint, it is desired but difficult to give a general specification that designates special knowledge to be extracted from texts of various special fields. In this paper, the specification table is given, for simplicity, for each specific field, and each kind of the main subject of technical papers though the scheme of the specification tables is almost the same.

Each sentence of the text is parsed semiautomatically with precise syntactic and general semantic information. If certain serious dependency ambiguities of terms are found that cannot be resolved by the ordinary category reference, the specific knowledge of the subframe associated with the term is retrieved from a knowledge database to resolve them. If the ambiguity cannot be resolved yet, then the system asks the user the correct dependency relation among terms.

After the parsing, the internal expression is normalized into a form similar to the specification. The specified information is extracted by scanning the normalized internal expressions of the text several times and stored in a form of relational tables or modified inverted files.

The case labels and the category names appearing in the internal expressions of technical sentences as well as in the subframes of specification tables were shared in both English and Japanese. Thereby, the extracted information can be immediately represented by the terms of a specified language by means of term-by-term replacement without serious deviation of meaning.

SPECIFICATION FOR EXTRACTED INFORMATION

The information to be extracted is designated by a specification table. It consists of several subframes related to the subject term:

$$L: \ (K_1\text{-}C_1:\_, \ldots, K_i\text{-}C_i:\underline{t}, \ldots, K_n\text{-}C_n:\_) \qquad (1)$$

0095-2338/84/1624-0015$01.50/0