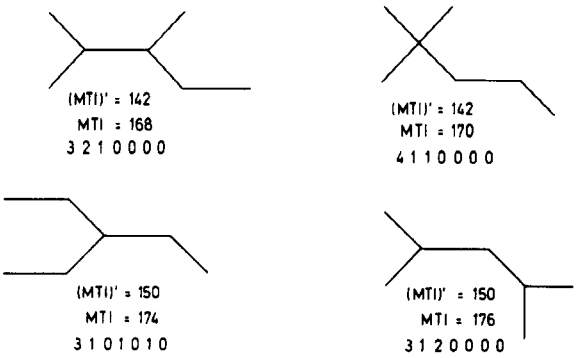

Figure 6. Two pairs of heptane trees with degenerate (MTI)′ values. Beneath each tree its MTI number and *N*-tuple code are also given.

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

Finally, we considered the row matrices $v(A + D)$ as possible *N*-tuple descriptors. The results of the analysis, which was limited to 18 030 alkane trees and several thousand cyclic structures, was that the row matrices are unique for the set of graphs investigated.

ACKNOWLEDGMENT

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

REFERENCES AND NOTES

(1) Schultz, H. P. Topological Organic Chemistry. 1. Graph Theory and Topological Indices of Alkanes. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 227–228.
(2) Knop, J. V.; Müller, W. R.; Jeričević, Ž.; Trinajstić, N. Computer Enumeration and Generation of Trees and Rooted Trees. *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 91–99.
(3) Trinajstić, N.; Jeričević, Ž.; Knop, J. V.; Müller, W. R.; Szymanski, K. Computer Generation of Isomeric Structures. *Pure Appl. Chem.* **1983**, *55*, 379–390.
(4) Knop, J. V.; Müller, W. R.; Szymanski, K.; Trinajstić, N. *Computer Generation of Certain Classes of Molecules*; SKTH/Kemija u industriji: Zagreb, 1985.
(5) Müller, W. R.; Szymanski, K.; Knop, J. V.; Trinajstić, N. An Algorithm for Construction of the Molecular Distance Matrix. *J. Comput. Chem.* **1987**, *8*, 170–173.
(6) Randić, M. On the Characterization of Molecular Branching. *J. Am. Chem. Soc.* **1975**, *97*, 6609–6615.
(7) Kier, L. B.; Hall, L. H. Molecular Connectivity in Chemistry and Drug Research; Academic: New York, 1976.
(8) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure–Activity Analysis*; Wiley: New York, 1986.
(9) Rouvray, D. H. The Limits of Applicability of Topological Indices. *J. Mol. Struct. (THEOCHEM)* **1989**, *185*, 187–201.
(10) Trinajstić, N. *Chemical Graph Theory*; CRC: Boca Raton, FL, 1983; Vol. 1, Chapter 4.
(11) Barysz, M.; Jashari, G.; Lall, R. S.; Srivastava, V. K.; Trinajstić, N. On the Distance Matrix of Molecules Containing Heteroatoms. In *Chemical Applications of Topology and Graph Theory*; King, R. B., Ed.; Elsevier: Amsterdam, 1983; pp 222–230.
(12) Randić, M. Compact Molecular Codes. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 136–148.
(13) Randić, M.; Nikolić, S.; Trinajstić, N. Compact Molecular Codes for Polycyclic Systems. *J. Mol. Struct. (THEOCHEM)* **1988**, *165*, 213–228.
(14) Bonchev, D.; Trinajstić, N. Information Theory, Distance Matrix and Molecular Branching. *J. Chem. Phys.* **1977**, *67*, 4517–4533.

Extraction of Chemical Reaction Information from Primary Journal Text

C. S. AI, P. E. BLOWER, JR.,* and R. H. LEDWITH
Chemical Abstracts Service, Columbus, Ohio 43210

Received November 28, 1989

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

INTRODUCTION

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

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

0095-2338/90/1630-0163$02.50/0

a **2,6-Bis[3-(bromomethyl)-2-methoxy-5-methyl-phenyl]-4-phenylpyridine (12e).** To a solution of **12d** (2.0 g, 4.4 mmol) in 20 mL of benzene was slowly added phosphorous tribromide (0.45 mL, 4.8 mmol) at 5 °C, whereupon the reaction mixture was stirred for 16 h at room temperature. After the addition of 50 mL of water, the mixture was neutralized with 10% $NaCO_3$. After the layers were separated, the aqueous phase was extracted with chloroform (2 × 50 mL). The combined organic phases were dried ($MgSO_4$), and the solvent was evaporated in vacuo to give **12e** as a white foam: yield 82%; mass spectrum, *m/e* 579.035 (M^+, calcd 579.041); ^1H NMR δ 8.06 (s, 2 H, pyridine H), 7.73-7.26 (m, 9 H, Ar H), 4.64 (s, 4 H, CH_2), 3.63 (s, 6 H, OCH_3), 2.69 (s, 6 H, CH_3).

Anal. Calcd for $C_{29}H_{27}Br_2NO_2$: C, 59.91; H, 4.68; N, 2.41. Found: C, 59.88; H, 4.74; N, 2.20.

b

Product:	2,6-Bis[3-(bromomethyl)-2-methoxy-5-methylphenyl]-4-phenylpyridine (12e)	**Yield:** 82%
Step: 1.1	**Event:** COMBINE	**Temp:** 5°C
Role	**Substance**	**Amount**
Reactant	12d	2.0 g, 4.4 mmol
Solvent	phosphorous tribromide	0.45 mL, 4.8 mmol
Solvent	benzene	20 mL
Step: 1.2	**Event:** REACT	**Temp:** room
	Time: 16 hr	

Figure 1. (a) Sample synthetic paragraph. (b) Synthesis frame.

numerical data are ignored because it is too expensive to enter or verify manually. The work described in this paper is potentially important in assisting to build a more comprehensive CASREACT database.

Using Zamora's work as a model, we have developed a series of programs that generate a summary of all preparative reactions from the experimental section of a JOC paper. This summary is represented as a frame,[3,4] called the *synthesis frame*. It identifies each participating substance, with its reaction role and quantity, and reaction times and temperatures. The synthesis frame also records some procedural information such as the order of mixing the reactants, reagents, and solvents and the duration and temperature of individual steps. For multistep reaction sequences, it identifies the generic class names of intermediate products. But the synthesis frame does not record workup procedures such as methods for isolating and purifying the product or data corroborating structural assignments. Figure 1a[5] shows a typical description of a synthetic preparation that we will use to illustrate the processing, and Figure 1b shows the synthesis frame created for this paragraph.

Processing Stages. The processing is divided into four stages. The first stage preprocesses the primary journal file. This involves a series of programs to locate papers on the primary journal tapes; extract the experimental section; translate the text to an ASCII representation that captures font information; mark word, sentence, and paragraph boundaries; and build a LISP data structure for the text. The second stage performs preliminary word classification mainly by dictionary lookup and word morphology. For example, word morphology rules are used for recognizing chemical line formulas that do not appear in the program's dictionaries. We also use suffix-stemming rules for recognizing adjectives, adverbs, and various words derived from verbs. This is similar to the processing performed by Zamora's program.[1]

The third stage (semantic processing) transforms the text of an experimental paragraph into a sequence of frames representing the meaning. Subprocesses search each sentence for important phrases (e.g., substance information, reaction conditions), split compound sentences into simple sentences, and convert each simple sentence into a frame. The final stage is a rule-based system[3,4] for building the synthesis frame. This program reads the sequence of frames that represent the meaning of each sentence, assigns roles to substances, eliminates sentences that do not contain important synthetic information (e.g., sentences describing the workup procedure), and builds the synthesis frame. As part of this processing, we have developed techniques for handling complex paragraphs like general procedures, analogous syntheses, etc. Since it deals with interparagraph (or *anaphoric*) references, this process has to treat the experimental section of the document as a whole.

In the remainder of this paper, we will focus on the processing done in the last two stages. The input to the semantic processing is the text of the entire experimental section of a document. The processing that precedes this stage has marked each word[6] (including numbers, punctuation, etc.) according to its syntactic class and may have added other important properties such as the root form of a verb and the Registry Number of a known substance. In addition, the incoming data also contain markers signaling the beginning of each new paragraph and sentence.

IDENTIFYING EVENTS

The primary purpose of the semantic stage is to convert the text of a paragraph into a series of frames describing the sequence of elementary events in the synthesis. In its simplest form, this sequences of events[7] is

combine reactants/reagents
allow reaction to proceed under stated conditions
quench reaction
isolate product
purify product
report yield
report characterization data

and the frame representation is intended to mimic this scenario.

This semantic processing is done in two phases. Phase I searches each sentence for important phrases (e.g., actions, substance information, reaction conditions). Then it splits compound sentences into independent clauses or simple sentences, each of which contains a verb. Each clause corresponds to a single event in the synthesis, which is converted in phase II to an *event frame* (or just *event*). The event frame is based on the verb, and each verb is mapped into one of eight primitive types. Thus, in the semantic processing, each sentence gives rise to one or more frames that mimic the meaning of the sentence but in a simplified and canonical form. At the conclusion of this stage, all subsequent processing is based on the event frames. Figure 2 illustrates this processing for the first sentence of the sample paragraph in Figure 1a.

Phase I. Phase I searches each sentence for important sentence fragments describing actions, substances, and reaction conditions. This search involves a partial parse of the sentence using Augmented Transition Network (ATN) parsing[8-10] to match templates against sentence fragments. We do not attempt to parse each sentence as a whole because a full parse of the sentence appears to be unnecessary on the basis of our experience. This avoids the complexities normally associated with natural language processing.

The search (or partial parse) proceeds from left to right through the sentence, one word at a time. At each word, the series of ATN templates is matched against the remainder of the sentence. The templates are prototype phrases describing (1) substance information, (2) references to external procedures, (3) time/temperature data, (4) verb phrases, and (5) characterization data.

If the ATN matching is successful, two items are returned: (1) the matched phrase encapsulated in a frame representation and (2) the unprocessed portion of the sentence. The text of the matched phrase is then replaced by the frame, and processing continues with the unprocessed portion of the sentence.

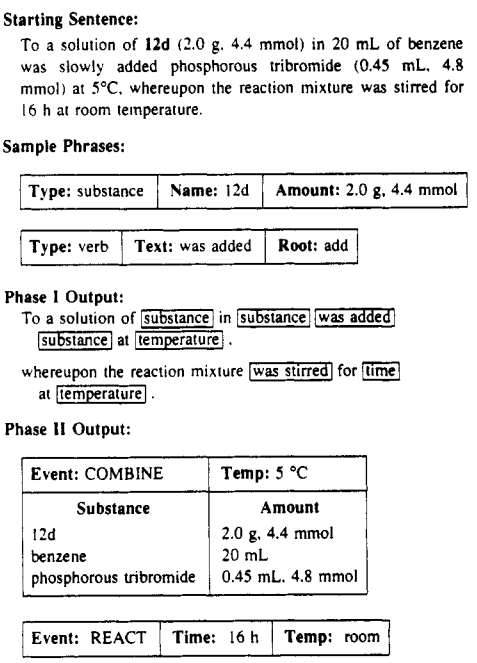

Figure 2. Semantic processing.

When all phrases have been identified, the procedure splits the sentence into a series of clauses, each containing one verb phrase.

The top portion of Figure 2 gives an example illustrating the processing done in phase I. The original sentence is at the top, followed by a frame for a substance phrase and a frame for a verb phrase. The output from phase I (in conceptual form) is shown next. The encapsulated phrases resulting from ATN matching are shown as boxed items which contain internal details that are not shown. The sentence in this example was divided into two clauses.

Handling of Verbs. A major objective of the semantic processing is to divide the experimental paragraph into a sequence of elementary events. Since verbs are the key to these events, we developed an ATN to recognize verb phrases. In full English writing and discourse, verb phrases can be quite complex.[11] Fortunately, we could avoid most of this complexity by taking advantage of the limited subject area we are dealing with and the narrative style used by technical authors.

In experimental paragraphs, sentences are written in the passive voice and past tense. The basic pattern for the verb phrases we are looking for is

(AUX{1,2} ADVERB?)? VERB PREP?

The pattern in parentheses is an optional phrase consisting of one or two auxiliaries followed by an optional adverb. This phrase is followed by the verb. In addition to simple verbs in the past tense, we also look for constructions like *allowed to* ⟨verb⟩ or *used to* ⟨verb⟩ and treat them as derivatives of whatever ⟨verb⟩ is. Finally, there might be a trailing preposition as part of a multiword verb like *take up* as in *the mesylate was taken up in ethanol.*

Not all verbs in a sentence are active (i.e., functioning as a verb); they can serve a number of other roles. The present particle form of a verb (i.e., ⟨verb⟩+*ing*) is particularly good for illustrating some of the different roles that a verb can fill (see Table I). Note that present participles do not occur as such in the documents we are dealing with. In fact, with two exceptions discussed below, the Verb-ATN will not identify them as active verbs.

The Verb-ATN is not dealing with the sentence as a whole; rather it is only matched against small, localized word groups as the sentence is processed from left to right. First, it is trying to identify active verbs taken, to some extent, out of context.

Table I. Syntactic Roles of Present Participles

role	sample phrase
object of preposition	with subsequent *stirring* for 1 h
preposition	mesylate was obtained *following* the general procedure
adverbial clause	after *being* washed with ether
adjective	the *coupling* reaction
gerund	*stirring* was continued for 30 min

The initial verb phrase in a sentence usually contains the auxiliary *was*, which clearly marks the following verb as active. But this is not always true of the remaining verbs in a compound sentence. Here is a typical example:

> After 4 h at room temperature, the mixture was hydrolyzed with ice–water, acidified with 30% H_3PO_4, and extracted with hexane.

Second, it tries to identify phrases containing verbs but in which the verb is functioning in another role. These phrases are shown by the patterns and examples

pattern	examples
(ART|PREP) (ADVERB? VERB,?)+	a freshly distilled under reduced pressure
until ANY+ VERBPAST	until the color disappeared
by (being VERBPAST|verbING)	by being heated by refluxing

There are two important situations for which we wish to treat a phrase as though it contained an active verb, even though that is not syntactically correct.

Yield words are important because they signal the arrival of a reaction product or intermediate. They have to be treated as special cases for two reasons: (1) They occur in constructs that are not otherwise recognized as verbs. (2) The main verb often refers to a workup action, and generally workup sentences are ignored. The following example is typical:

> The product was recrystallized from diethyl ether to afford **14c** as white crystals...

The patterns ⟨yield-word⟩+*ing* and *to* ⟨yield-word⟩ are marked as active verbs where ⟨yield-word⟩ is one of {*yield*, *give*, *provide*, *afford*, *furnish*}.

After-Clauses. There are several common constructions in which sentences begin *After* ⟨*event-1*⟩, ⟨*event-2*⟩. We would like to separate the two events, but the after-clause may not contain a verb phrase. So these constructs are also treated as special cases. Three different patterns are recognized:

pattern	example
After ⟨time⟩	After 3 h at 0 °C...
After ⟨verb⟩+ing	After cooling the mixture at 0 °C...
After ⟨verb⟩+tion	After ether extraction...

In the first example above, the implicit verb *react* is added to the clause. In the second and third examples, the verb is simply marked as active. Note that other adverbial clauses are not treated in this way. In particular, until-clauses are superficially analogous to after-clauses. However, an until-clause is not a separate event but an alternative (nonquantitative) means of expressing a time period; for example

> Ozone was bubbled through the alkene **10** ... until a pale blue color persisted.

Phase II. Phase II converts the series of clauses produced by phase I into event frames illustrated in the bottom portion of Figure 2 for the sample sentence. Each clause is processed

Table II. Primitive Acts Used in Event Frames

type	description of verbs mapped to type
COMBINE	verbs indicating that substances were combined; e.g., add, dissolve, introduce, treat
REACT	verbs indicating that the reaction was proceeding; e.g., heat, react, reflux, stir
PREPARE	verbs pointing to the product and possibly a reactant; e.g., convert, generate, make, prepare, synthesize
WORKUP	verbs indicating that the reaction was stopped or referring to isolation or purification of the product; e.g., collect, concentrate, extract, filter, purify
RESULT	verbs indicating that the product was obtained; e.g., give, obtain, provide, yield
TITLE	no verb; used for paragraph titles
MISC	verbs suggesting that the sentence is probably not important; e.g., act, agree, appear, apply
UNKNOWN	*no verb recognized or no default meaning*

independently, and the frame-building procedure is a simple loop that moves through the clause looking for the phrases identified in phase I. The focus of the processing is the verb phrase, and generally each clause contains one verb phrase. The handling of verbs involves some special rules, but all other phrases—substance, time, temperature, procedure, and yield phrases—are simply accumulated. Table II gives the eight primitive acts that are used as frame types and indicates the types of verbs that are mapped to each.

CREATION OF THE SYNTHESIS FRAME

This is the last stage in the processing flow. It extracts and organizes reaction information and creates a synthesis frame for each preparation described in the experimental section. The input to this program is a sequence of unorganized event frames. The program first removes paragraphs that do not describe preparations. Then it creates a synthesis frame for each paragraph[12] and organizes them into a tree structure. Finally, it processes any complex procedures that involve interparagraph references.

Organization and General Processing. The processing that precedes this stage has converted the text of each paragraph into a set of events. Each event is one of those listed in Table II. The important reaction information appears in PREPARE, RESULT, REACT, or COMBINE events. In addition, a TITLE event is useful in determining the final product and handling interparagraph references. Information in a WORKUP, UNKNOWN, or MISC event is superfluous (at least for our purposes) and is simply removed.

To facilitate the implementation of interparagraph references, the input is organized into a tree structure as shown in Figure 3. The tree corresponds to the complete experimental section of a journal paper, except paragraphs that are not preparative. An experiment consists of one or more paragraphs, headed by a title and containing synthesis information. A title usually identifies the final products and sometimes also the reactants. Some paragraphs may be general procedures and contain a TYPE property. The synthesis information consists of one or more steps which in turn contain reaction information encapsulated in the sequence of event frames. The next two sections give the rules for assigning roles to substances and the rules for converting the stream of event frames into a synthesis frame.

Assigning Roles to Substances. Each substance in a synthesis frame is assigned one of the following roles: REACTANT, PRODUCT, REAGENT, SOLVENT, or CATALYST. Rules for assigning these roles are given in the following list. In addition to the five main roles, there are also subcategories in some cases (rules 8–11).

(1) If a substance is known (i.e., if it was found in the dictionary), then use its default role.

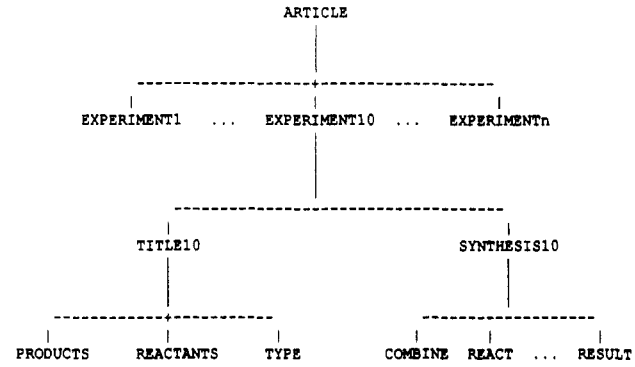

Figure 3. Organization of experimental section.

(2) If a substance is marked as generic,[13] then it is a product or reactant.

(3) If a substance has an ID (e.g., **12e** in Figure 1a), then it is a product or reactant.

(4) If a product or reactant is in a COMBINE event, then it is a reactant.

(5) If a substance has a VOLUME property, then it is a solvent.

(6) If a substance is represented as a chemical formula, then it is a reagent.

(7) The default role is reagent. Note that at this time we have no rules for recognizing catalysts.

(8) Byproducts. The last event of a paragraph may contain several products; all except the title product are marked as byproducts.

(9) Intermediates. Any products other than the ones in the last event are marked as intermediates.

(10) Unreacted starting material. If a product of the last event matches one of the reactants mentioned earlier, then it is marked as unreacted starting material.

(11) Unspecified substances. If a generic substance is copied from a general procedure, then it is marked as an unspecified substance.

Conversion Rules. The following list gives some of the general rules for converting the stream of event frames to a synthesis frame. Other techniques or rules are described under Complex Paragraphs.

(1) Remove paragraphs that contain no reaction information. Either the title indicates that the paragraph is not a preparation or the paragraph contains only data.

(2) Remove any COMBINE event that appears in a WORKUP–COMBINE–WORKUP or a REACT–COMBINE–WORKUP sequence. This heuristic is based on the fact that a substance added in one of these situations is usually part of the workup procedure.

(3) Remove WORKUP, UNKNOWN, and MISC events from the input.

(4) If a substance has an unknown ID, then append the rest of the information to the previous substance and remove the substance.

(5) If a RESULT event only has YIELD property, then attach the YIELD property to the previous product and remove the RESULT event.

(6) If there are two consecutive COMBINEs that do not have different TEMP properties, then merge them and add the quantities of any common substances.

(7) If a REACT event follows a COMBINE event and they do not have different TEMP properties, then move substances in the REACT event into the COMBINE event. If there are common substances, add the quantities of the common substances.

(8) If the product of an intermediate RESULT event is not mentioned in the subsequent COMBINE event, then copy it into the COMBINE event and change its role to REACTANT. Heu-

a **Synthesis of Precursors. General Procedure for the Preparation of the Nitriles.** The appropriate alcohol (4 mmol) was dissolved in distilled methylene chloride (20 mL) and triethylamine (1 mL), followed by methanesulfonyl chloride (0.45 mL) and catalytic amounts of 4-(dimethylamino)pyridine (DMAP, 20 mg) added dropwise at 0° C. The reaction was stirred at room temperature for 8 h. Aqueous $NaHCO_3$ (5%) was added and the reaction mixture was extracted with CH_2Cl_2. The organic layer was dried over sodium sulfate and evaporated, and the crude mesylate was purified by column chromatography on silica gel (yield 80%). The methanesulfonate (0.34 mmol) was dissolved in 12 mL of tetrahydrofuran/dimethyl sulfoxide (1:1 v/v) and $K^{14}CN$ (1.3 mg, specific activity 57.6 mCi/mmol) and cold KCN (15 mg) were added. The reaction mixture was refluxed for 8 h under nitrogen. After the usual workup with ether (4 × 15 mL), the desired nitrile was obtained and purified by silica gel column chromatography with hexane/ether (4:1 v/v) as eluent (yield 85%).

b

Product:	Nitriles	**Type:** GP/Series
		Yield: 85%
Step: 1.1	**Event:** COMBINE	**Temp:** 0°C
Role	**Substance**	**Amount**
Reactant	non-spec alcohol	4mmol
Solvent	methylene chloride	20 mL
Solvent	trimethylamine	1 mL
Solvent	methanesulfonyl chloride	0.45 mL
Reagent	4-(dimethylamino)pyridine	20 mg
Step: 1.2	**Event:** REACT	**Temp:** room
	Time: 8 hr	
Step: 1.3	**Event:** RESULT	
	Intermediate: mesylate	**Amount:** 80%
Step: 2.1	**Event:** COMBINE	
Role	**Substance**	**Amount**
Reagent	methanesulfonate	
Solvent	tetrahydrofuran/ dimethyl sulfoxide	12 mL 1:1
Reagent	$K^{14}CN$	1.3 mg
Reagent	KCN	15 mg
Reagent	nitrogen	
Step: 2.2	**Event:** REACT	**Time:** 8 h

Figure 4. (a) General procedure. (b) Synthesis frame.

ristic: The product of an intermediate step is a reactant in the following step.

(9) Group events in a synthesis list into steps; each intermediate step contains a list of events terminating with a RESULT event.

Complex Paragraphs. We can distinguish several different types of complex paragraphs which we discuss individually below. Most involve some sort of interparagraph reference.

General Procedures. Parts a[14] and b of Figure 4 illustrate a common method for reporting a series of similar reactions. The general procedure is a condensed format which allows the author to avoid repeating those reaction conditions that remain constant. The text is basically a template containing some substances or conditions that are specific and others that are generic or variable. Variable substances are referred to by such phrases as *the appropriate alcohol* and *the desired nitrile*. The specific substances actually involved will be reported elsewhere in the experimental section.

There are two pieces of information that are important and need to be recorded in a general procedure. First, a paragraph must be recognized as a general procedure. Usually, the title will mention *general procedure*, and this information is recorded as the TYPE property. Second, all generic substances in a general procedure must be identified and marked; when references to the general procedure are processed, these substances can either be replaced with specific substances or marked as ambiguous.

Instances of a General Procedure. These procedures, illustrated by parts a[14] and b of Figure 5, are skeletal and often only indicate the specific substances that are to be substituted

a **dl-1-[^{14}C]Cyano-9-methylpentadecane** was obtained following the general procedure and further purified by reverse phase HPLC (Altex, MeOH; yield 86%): specific activity = 4.22 mCi/mmol; ^{1}H NMR (300 MHz, $CDCl_3$) δ 0.833 (3 H, d, J = 6.0 HZ, 9-CH_3), 0.880 (3 H, t, J = 6.3 HZ, RCH_3), 2.332 (2 H, t, J = 7.2 HZ, RCH_2CN); MS (70eV), *m/z* (relative intensity)251 (M^+, 2.8), 235 (4.7), 222 (6.5), 208 (8.4), 194 (8.4), 180 (9.3), 166 (31.8), 152 (10.3), 138 (19.6), 124 (23.4), 110 (30.8), 96 (34.6), 82 (32.7), 71 (53.3), 57 (82.2), 41 (100).

b

Product:	dl-1-[^{14}C]Cyano-9-methylpentadecane	**Yield:** 86%,
Step: 1.1	**Event:** COMBINE	**Temp:** 0°C
Role	**Substance**	**Amount**
Reactant	non-spec alcohol	4mmol
Solvent	methylene chloride	20 mL
Solvent	trimethylamine	1 mL
Solvent	methanesulfonyl chloride	0.45 mL
Reagent	4-(dimethylamino)pyridine	20 mg
Step: 1.2	**Event:** REACT	**Temp:** room
	Time: 8 hr	
Step: 1.3	**Event:** RESULT	
	Intermediate: mesylate	**Amount:** 80%
Step: 2.1	**Event:** COMBINE	
Role	**Substance**	**Amount**
Reagent	methanesulfonate	
Solvent	tetrahydrofuran/ dimethyl sulfoxide	12 mL 1:1
Reagent	$K^{14}CN$	1.3 mg
Reagent	KCN	15 mg
Reagent	nitrogen	
Step: 2.2	**Event:** REACT	**Time:** 8 h

Figure 5. (a) Instance of general procedure. (b) Synthesis frame.

for generic substances (variables) in a general template. The processing done for this type of paragraph follows a scenario that is also used for other types of interparagraph references. It consists of three steps: (1) locate the referenced procedure, (2) copy the corresponding synthesis frame, and (3) replace substances in the referenced paragraph with substances mentioned in the referencing paragraph.

In this instance, the program first attempts to locate the referenced paragraph using the product in the referencing paragraph by comparing it with all products of general procedures in the document. Generic and specific products are considered to match if either the name of the generic product is contained in the specific product or it is a synonym. In this case, *cyano* is a synonym of *nitrile*, so the general procedure for nitriles (Figure 4a) is assumed to be the referenced paragraph. Then the program copies the corresponding synthesis frame and examines the generic substances. All instances of *nitrile* are replaced with the product from the referencing paragraph. Finally, since there is no replacement for the nonspecific alcohol, it remains ambiguous.[15]

Analogous Syntheses. Figure 6a–c[5] illustrates syntheses that are described as being analogous or similar to another procedure. Again these procedures are skeletal, much like the instances of a general procedure. But there is an additional problem in locating the variables in the referenced procedure since they will not be generic substances. The program handles this paragraph following the scenario used for general procedures. It uses the information in the REF slot (i.e., ID 12e in Figure 6b) as the search term and compares that to each of the products in the document. If a matching product is found, then its synthesis frame is copied. Then the program attempts to find replacements for each substance using the following rules: (1) Replace substances with matching ROLES. (2) Replace substances with like properties, e.g., a substance with an ID should replace one with the same property. Finally, if no quantity is given for the replacing substance, then normalized quantities (e.g., millimole) are copied from the re-

a **2,6-Bis[3-(bromomethyl)-2-methoxy-5-methyl-phenyl]-pyridine (13e)** was obtained from **13d** as described for **12e**. The product was recrystallized from diethyl ether to give colorless crystals: yield 82%; mp 105-106 °C; mass spectrum, *m/e* 503.007 (M^+, calcd 503.010); ^1H NMR δ 7.79 (s, 2 H, pyridine H), 7.60 (d, 2 H, Ar H), 7.24 (d, 2 H, Ar H), 4.62 (s, 4 H, CH_2), 3.58 (s, 6 H, OCH_3), 2.37 (s, 6 H, CH_3).

Anal. Calcd for $C_{23}H_{23}Br_2NO_2$: C, 54.68; H, 4.59; N, 2.77. Found: C, 54.85; H, 4.70; N, 2.57.

b

Product:	2,6-Bis[3-(bromomethyl)-2-methoxy-5-methylphenyl]-pyridine (13e)		**Yield:** 82%
Event:	PREPARE	**Reactant:** 13d	**Refer:** 12e

c

Product:	2,6-Bis[3-(bromomethyl)-2-methoxy-5-methylphenyl]-pyridine (13e)	**Yield:** 82%
Step: 1.1	**Event:** COMBINE	**Temp:** 5°C
Role	**Substance**	**Amount**
Reactant	13d	4.4 mmol
Solvent	phosphorous tribromide	0.45 mL, 4.8 mmol
Solvent	benzene	20 mL
Step: 1.2	**Event:** REACT	**Temp:** room
	Time: 16 hr	

Figure 6. (a) Analogous procedure. (b) Initial synthesis frame. (c) Synthesis frame.

placed substance, but weights (e.g., milligram) are not.

Parallel Syntheses. Another technique authors use to condense the description of several similar preparations is to describe them as though they were a single reaction. Figure 7a–c[5] shows a situation in which the syntheses of several substances are described in parallel. The first paragraph gives a general procedure for preparing substances **8a**, **8b**, and **8c**, and the following paragraphs give details for each one. The clues used to recognize this situation are the plural product name and the substance IDs written as a series. One difficulty here is that the referenced paragraph is not mentioned explicitly in the referencing paragraph. But since **8a** is one of the products listed in the general procedure, the synthesis frame for it is copied for **8a**. Then all products except **8a** are deleted from the list containing **8a**, and all reactants except the corresponding **5b** are deleted from the list containing **5b**. Figure 7b shows the general synthesis frame created from the first paragraph of Figure 7a, and Figure 7c shows the synthesis frame created from the general frame and the second paragraph.

FUTURE RESEARCH

A long-range goal of this research is to lay the groundwork for a program that could take over most of the work now done by document analysts to generate CASREACT records from ACS manuscripts. We have not yet achieved that goal. As it now stands, the program developed in this project is not robust or reliable enough to be used in production. For simple synthesis paragraphs, the program produces usable results in the range of 80–90%, but for complex paragraphs this drops to 60–70%.

We have not, however, discovered any major obstacles in using the processing strategy described here for extracting reaction data from the experimental section of a journal paper. But further research is needed to refine or revise these techniques before work on a production program can be initiated. First, we would expand the dictionaries so that the program could recognize virtually every word. The only words the program should not be expected to recognize are new chemical substances and author names. A notable deficiency in the current system is the lack of a dictionary of common chemical names. Our dictionaries do contain common acronyms and

a **General Procedure for the Preparation of the Pyrylium Tetrafluoroborates 8a-c.** To a mixture of the ethanone **5b** (**7a**, **7b**) (0.2 mol) and freshly distilled benzaldehyde (10.6 g, 0.1 mol) was added boron trifluoride etherate (38.6 g, 0.2 mol) at 70 °C. The reaction mixture was heated for 2 h at 70 °C meanwhile allowing diethyl ether to evaporate from the reaction mixture. After cooling to room temperature, the pyrylium salt was filtered off or was crystallized by addition of diethyl ether (250 mL) to the reaction mixture. After the crystals were washed with diethyl ether, the product was recrystallized (solvent). Characteristic ^1H and ^{13}C NMR data are given in Table I.

2,6-Bis(2-methoxy-5-methylphenyl)-4-phenylpyrylium tetrafluoroborate (8a) was obtained from **5b** as orange crystals (acetic acid): yield 20%; mp 225-230 °C; mass spectrum, *m/e* 397.180 (M^+, calcd for $C_{27}H_{25}O_3$ 397.182).

b

Product:	Pyrylium Tetrafluoro borates (8a, 8b, 8c)	**Type:** GP/Series
Step: 1.1	**Event:** COMBINE	**Temp:** 70 °C
Role	**Substance**	**Amount**
Reactant	ethanone (5b, 7a, 7b)	0.2 mol
Reagent	benzaldehyde	10.6 g, 0.1 mol
Reagent	boron trifluoride etherate	38.6 g, 0.2 mol
Reagent	diethyl ether	
Step: 1.2	**Event:** REACT	**Temp:** 70 °C
	Time: 2 hr	
Step: 1.3	**Event:** REACT	**Temp:** room

c

Product:	2,6-Bis(2-methoxy-5-methyl phenyl)-4-phenylpyrylium tetrafluoroborate (8a)	**Yield:** 20%
Step: 1.1	**Event:** COMBINE	**Temp:** 70 °C
Role	**Substance**	**Amount**
Reactant	ethanone (5b)	0.2 mol
Reagent	benzaldehyde	10.6 g, 0.1 mol
Reagent	boron trifluoride etherate	38.6 g, 0.2 mol
Reagent	diethyl ether	
Step: 1.2	**Event:** REACT	**Temp:** 70 °C
	Time: 2 hr	
Step: 1.3	**Event:** REACT	**Temp:** room

Figure 7. (a) Parallel syntheses. (b) General synthesis frame. (c) Specific synthesis frame.

line formulas but not common chemicals like acetone, ethyl acetate, pyridine, etc. Currently, these are recognized as chemical names because they contain the fragments *acet*, *ethy*, and *pyr*. But they should also be recognized as common substances, with Registry Numbers and probable reaction roles, in the same way that acronyms and line formulas are. We would also add multiword phrases (and semantic information) for such things as chemical apparatus and common laboratory procedures (column chromatography, X-ray crystallography, etc.).

Second, we would investigate a better parsing technique. The ATN parsing currently used has been the source of numerous problems. It should be replaced by a method that is more robust and easily extendible. Third, we would implement a true rule-based system for use in the creation of the synthesis frames. The program does use heuristic rules for these functions, but currently they are embedded in LISP code. Recoding this using rule-based Expert System[4] technology would make the program easier to maintain and extend, even by a nonprogrammer. The final task is related and would be aimed at maturing the rule base. This can be achieved mainly through experience—processing many documents and gradually enhancing the program's performance.

REFERENCES AND NOTES

(1) Zamora, E. M.; Blower, P. E., Jr. Extraction of Chemical Reaction Information from Primary Journal Text Using Computational Linguistic Techniques. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 176, 181.

(2) STN is a registered trademark of the American Chemical Society.
(3) Barr, A.; Feigenbaum, E. A. *The Handbook of Artifical Intelligence*; Kaufmann: Los Altos, CA, 1981.
(4) Winston, P. H. *Artificial Intelligence*; Addison Wesley: Reading, MA, 1984.
(5) Dijkstra, P. J.; den Hertog, H. J.; van Steen, B. J.; Zijlstra, Z.; Skowronska-Ptasinska, M.; Reinhoudt, D. N.; van Eerden, J.; Harkema, S. *J. Org. Chem.* **1987**, *52*, 2433-42.
(6) A few words will not be recognized and are simply marked as unknown.
(7) Schank's school would call this a *script*; see: Schank, R. C.; Abelson, R. P. *Scripts, Plans, Goals and Understanding*; Erlbaum: Hillsdale, NJ, 1977. We could call it a *synthesis script*, but we have avoided this term, since we are not following Schank's methodology closely enough.
(8) Woods, W. A. Transition network grammars for natural language analysis. *Commun. ACM* **1970**, *13*, 591.
(9) Amsterdam, J. Augmented Transition Networks for Natural Language Parsing. *AI Expert* **1986**, *1*, 15-21.
(10) Charniak, E.; McDermott, D. *Artificial Intelligence*: Addison Wesley: Reading, MA, 1986.
(11) Quirk, R.; Greenbaum, S.; Leech, G.; Svartvik, J. *A Grammar of Contemporary English*; Longman: Harlow, England, 1972.
(12) The term *paragraph* will be used for a single paragraph or group of related paragraphs under a single title.
(13) A substance is marked as generic if it occurs with a definite article as in *the acid*.
(14) Raederstorff, D.; Shu, A. Y. L.; Thompson, J. E.; Djerassi, C. *J. Org. Chem.* **1987**, *52*, 2337-46.
(15) Resolving this type of ambiguity requires a sophisticated knowledge of chemical structure and reactivity and is beyond the scope of our current work.

User Needs in Chemical Information

GÜNTER PÖTZSCHER*
FIZ Chemie GmbH, Steinplatz 2, D-1000 Berlin 12, FRG

A. J. C. WILSON
Crystallographic Data Centre, University Chemical Laboratory, Lensfield Road, Cambridge CB2 1EW, U.K.

Received December 5, 1989

Information is of great value in our modern industrial society. In chemistry, a discipline in which compounds and compound classes play the most important role, information on about 10 million compounds has been registered. This number increases annually by 0.5 million compounds published in about 0.5 million documents. In this context it is a prerequisite that the information hidden in books, scientific and technical journals, conference proceedings, dissertations, etc. be evaluated and presented by abstracting and indexing services and database producers. Timeliness, accuracy, and completeness of the information are high on the list of desiderata of the users in chemistry. It is of great importance that all new information published in a primary source be considered and made searchable. Furthermore, properties of compounds including stereochemistry, toxicity, environmental behavior, etc. have to be made searchable, too. Factual and/or numerical data and reviews are requested as well. User friendliness of the services is of high priority. In the future, the electronic media will dominate the field of information more and more and therefore many improvements in search methods are necessary.

1. INTRODUCTION

Modern industrial society depends on the availability of the commodity "information". Information is of incalculable value to industry, government, and universities; its value far exceeds its cost, even though only the latter is readily expressed in monetary terms. The difficulty in putting a monetary estimate on its value explains to some extent the reluctance of government and industry to support adequately the necessary but "uneconomic" institutions that participate in the flow of information from producer to user. In scientific and technological fields the producer of information of one type is often the consumer of the same or of a different type.

1.1. Institutions Involved in Information Flow. The following institutions participate in the flow of information from producer to user:

1.1.1. Newspapers, Popular Journals, Radio, and Television. The popular media disseminate "news" rather than "information". It is often sensationalized, and even if it is reported seriously, it may be distorted by the failure of the reporter to understand it fully. For the chemist it can at most be an indication that fuller details must be sought elsewhere.

1.1.2. Commercial and Learned-Society Publications. Many commercial publishers and learned societies produce books and scientific and technical journals that contain original research, compilations, and reviews. Until recently such publications have been "hard copy" on paper, but some are now available in electronic form.

1.1.3. Patent Offices Publishing Patents and Related Works. Patents present special problems, as their purpose and mode of drafting differ completely from those of a journal paper. The purpose of the latter is to convey as much information as possible, and commercial considerations do not restrict the freedom of the author. An ideal patent, on the other hand, would contain no information or "know-how" that might be useful to a competitor, and the claims would be drafted so that all literature searches would lead to the patent. Patent law ordinarily prevents the patent draftsman from reaching this ideal, though some patents approach it.

1.1.4. Abstracting and Related Services. Abstracting and indexing services, data-bank compilers, handbook producers, and textbook authors analyze and report the original literature, thus making it more readily accessible. Although the activities of the organizations described here and in section 1.1.2 are conceptually distinct, the same parent organization often engages in activities of both types. For example, the American Chemical Society produces both the *Journal of the American Chemical Society* and *Chemical Abstracts*, and the International Union of Crystallography publishes two journals, an annual volume of abstracts, and the multivolume handbook *International Tables for Crystallography*.

1.1.5. Libraries. In principle, libraries should collect and store all types of literature, make them available to the user, and provide on-line access to electronic abstracting services and databases. In practice, however, underfunding restricts

0095-2338/90/1630-0169$02.50/0