Extraction of Chemical Reaction Information from Primary Journal Text Using Computational Linguistics Techniques. 1. Lexical and Syntactic Phases[†]

ELENA M. ZAMORA[‡] and PAUL E. BLOWER, JR.*

Chemical Abstracts Service, Columbus, Ohio 43210

Received December 30, 1983

An experimental technique for extracting specific facts about chemical reactions from the full text of American Chemical Society journals is described. The chemical information is used to encode Reaction Information Forms that can serve as the basis for a computer-readable file of chemical reactions. The programs use computational linguistics techniques to analyze the natural language text. The restricted semantic domain of the source publications makes it possible to resolve ambiguities effectively. This paper focuses on the lexical and syntactic aspects of the specialized information extraction procedures, and the following paper will describe the associated semantic considerations.

INTRODUCTION

The continuing growth in the number of publicly available databases and query languages is a clear indication of the importance of information in our daily lives. Yet, one of the major economic problems of building databases is the entry and verification of the original information into the databases. Advances in computational linguistics and the increased availability of large files of computer-readable natural language text generated as a byproduct of photocomposition offer the possibility of developing automatic database-building techniques. Currently, the most promising application of these techniques is in the extraction of information from the highly restricted semantic domain of specialized technical journals.[1,2]

For many years, technical secondary publications available in computerreadable form, such as *Chemical Abstracts*, have been sources for on-line databases provided by various information services. Typically, the publication title, bibliographic information, keyword entries, concept, author, and substance index entries are specially prepared to serve as access points to the publications. The text of the abstracts is also available in computer-readable form but has had limited use in on-line databases because of the large storage requirements. However, neither the fields used as access points nor the text of the abstracts contains the type of hard information necessary to build a file of chemical reaction data. The purpose of the index entries and the keywords is to provide a mechanism for locating documents relevant to a specific topic. The abstract provides enough information about the original document to allow the user to determine if the primary journal needs to be consulted. Therefore, primary journals, insofar as they are computerized, have the greatest potential as sources for the automatic generation of databases.

This paper describes techniques for extracting facts about chemical reactions from the text of primary journals of the American Chemical Society (ACS). In general, the descriptions of synthetic reactions that appear in the experimental sections of primary chemical journals have a fairly simple and predictable format and concern a very specific and well-defined subject, see Figure 1. The diversity of expression encountered in these paragraphs makes it necessary to use methods from computational linguistics to extract the chemical information.

The goal of the work described here is to generate Reaction Information Forms (RIFs) containing the reactants and reagents of a chemical reaction, along with their quantities (scale) and reaction conditions (including solvents, time, temperature, auxiliary energy sources, apparatus, etc.). Figure 2 contains an example of an RIF produced manually from the synthetic description given in Figure 1. Storing the chemical reaction data in the form of a table facilitates the retrieval of specific facts and verification of whether desired relations exist between the textual or numeric fields. Such a type of system was developed by the Linguistic String Project at New York University.[3] The desire to automatically generate the RIF originated from the fact that manual encoding is extremely tedious and time consuming.

Paragraphs describing chemical reactions are subjected to two types of text analysis: a sentence analysis that serves to guide the identification of chemical substances within sentences and an extrasentential or discourse text structure analysis to take advantage of the format of the text. The sentence analysis uses a "word expert" approach[4,5] during the lexical and syntactic phases. The discourse structure analysis and a "frame" approach[6,7] are used to obtain a representation of the meaning of the paragraphs during the semantic phase.

Chemical substances, which comprise the reactants and the products of a chemical reaction, reaction conditions, and yield are identified by a hierarchical application of procedures. Multiple passes through the primary paragraph are made to analyze the sentences prior to the extraction of data. The initial passes consist of simple word lookup and morphological analysis of words and numeric strings (lexical phase). A secondary stage analyzes superficial sentence structure on the basis of fundamental parts of speech and punctuation (syntactic phase), and a third stage identifies sentences to be processes or skipped and provides a suitable meaningful representation for those to be processes (semantic phase). Finally, a detailed analysis of the sentences describing the synthesis is performed to extract information for completion of the RIF (extraction phase). The system outline is given in Figure 3.

INITIAL DATABASE

The current research is limited in scope to synthetic reactions from the *Journal of Organic Chemistry* (JOC), which follow a simple model. Some advantages of JOC are that the documents are already in computer-readable form and author-assigned substance numbers or names are linked to CAS Registry Numbers within the document. This journal provides over 600 papers per year containing chemical reaction information.

The "simple model" of a chemical reaction paragraph has been defined as one that describes the formation of only one product and has a discourse structure containing a heading, synthesis description, workup, and characterization of the product. The heading typically contains the chemical name of the reaction product and is not a complete sentence. The synthesis description consists of one or more sentences that list the reactants, their quantities, and reaction conditions such as time, temperature, and solvents. The workup describes the

[†] Presented in part before the Division of Chemical Information, 184th National Meeting of the American Chemical Society, Kansas City, MO, Sept 13, 1982.

[‡] Present address: P.O. Box 3554, Gaithersburg, MD 20878.

0095-2338/84/1624-0176$01.50/0

SIMPLE MODEL

Heading N-2-methyl-5,6-dihydro-7,10-dimethyl-1,4:4a,10b-diethenobenzo(f)phthalazine-2,3(1h,4h)-dicarboximide (7b).

Synthesis To 6b (2.42 g, 10.4 mol) dissolved in 50 ml of cold (−78+degree+c) 4:1 pentane/ethyl acetate was added dropwise N-methyltriazolinedione (1.17 g, 10.4 mmol) dissolved in ethyl acetate (14 ml). After the addition was finished, the reaction mixture was stirred for 1 h at room temperature.

Work-up The slightly pink solid was collected to give 2.38 g (67%) of *urazole 7b. The analytical sample was* prepared by recrystallization from benzene:

Characterization white solid; mp 220-222 +degree+c; ir (kbr)+nu+max 3100-2800, 1770, 1705, 1450, 1390, 1380, 1190, 850, 800, 780, 740, 605 cm-1; 1H NMR (CDC13)+delta+6.97 (s, 2 H), 6.4-6.1 (m, 3 H), 6.12 (d, J = 2.9 Hz, 1 H), 5.10-4.75 (m, 2 H), 3.10-2.75 (m, 2 H), 2.95 (s, 3 H), 2.5-1.9 (m, 2 H), 2.35 (s, 3 H), 2.23 (s, 3 H);13C NMR (CDC13) 158.8 (s), 158.4 (s), 143.4 (d), 139.5 (d), 137.7 (s), 134.1 (s), 133.6 (s), 133.1 (s), 129.1 (d), 128.9 (d), 128.7 (d), 128.2 (d), 60.7 (2C, 2 d), 50.8 (s), 46.8 (s), 28.5 (t), 26.1 (t), 25.3 (q), 20.6 (q), 20.3 ppm (q); mass spectrum, calcd m/e 347.1634, obsd 347.1642.

Anal. Calcd for C21H21N302: C, 72.60; H, 6.09. Found: C, 72.71; H, 6.17

Figure 1. Reaction description, divided to show components of the simple model.

REF	SCALE small	PHASE liquid	YIELD 67%	TEMP. −78 to 20
TIME 1 h	ENERGY	APPARATUS	FEATURES: NMR, IR, MS	
REG. NO.	FUNCTION	AMOUNT	AUTHOR ID.	
78624-62-1	product	2.38 g	7b	
78624-61-0	reactant	2.42 g	6b	
13274-43-6	reactant	1.17 g	N-methyltriazolinedione	
	solvent	40 mL	pentane	
	solvent	24 mL	ethyl acetate	

Figure 2. *Reaction Information Forms, produced manually from the descriptions of Figure 1.*

procedure for completing the reaction and for isolating and purifying the product. Near the end of the workup, the yield is stated. Finally, the characterization provides analytical data that can be used to confirm the identity of the product.

Although these constraints would seem to be restrictive, over half of a randomly selected sample of 321 paragraphs followed the simple model. The highly structured manner in which JOC descriptions are organized facilitates manipulation of the discourse structure but does not result in an oversimplification of the problems. The procedures used to identify the components of a reaction within the sentences depend only on the syntactic structure and are likely to be the same whether the descriptions conform to the simple model or not, but the semantic problems are much more complex for the more general cases.

The simple model is based on a purely statistical phenomenon. Neither the authors nor the editors of the journal make a decision to organize a paragraph according to the simple model. Rather, the simple model is a logical way of describing a reaction, and it is inevitable that this type of organization will be used. However, considerations of style or space sometimes result in synthetic paragraphs that require substantial chemical knowledge and text-scanning capability for resolution. A sentence such as "The procedure is as for 5a above and with either 9 or 10 gives 9a." requires finding a previously described procedure for synthesizing substance 5a and subsequently substituting the starting material in the previous description with substance 9 or substance 10 to find out how to produce substance 9a. Paragraphs containing such anaphoric references or dependencies on tables or structural drawings were not selected for this study.

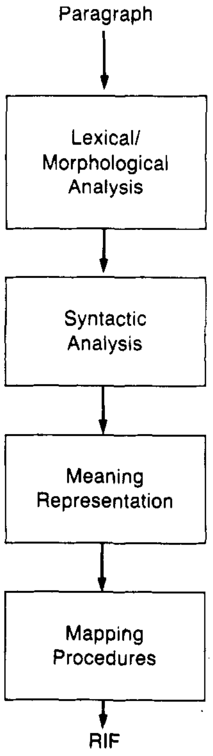

Figure 3. System components creating the Reaction Information Form.

The programs were designed to work on paragraphs from the experimental sections of JOC that were obtained from the ACS on-line primary journal files available through Bibliographic Retrieval Service (BRS).[8] The photocomposition files use a large character set, but the on-line files have a restricted character set with no boldface characters, superscripts, italics, or upper- and lower-case character distinctions. One advantage of this uniformity is that dictionary lookup and some aspects of morphological analysis are simplified, but it is not possible to distinguish between the abbreviation for hour (h) and the symbol for hydrogen (H) except from the context. Another feature of the BRS data is that some characters, such as degree symbols and Greek letters, are represented as multiletter strings (e.g., +degree+, +alpha+, etc.). Use of this character set has not been detrimental to our study, but improvements resulting from the use of a full character set are a subject for further study.

LEXICAL PHASE

The lexical phase, as its name implies, is fundamentally dependent on the use of dictionaries (a lexicon) to classify words of the text into categories. This phase involves (1) isolating the words in the paragraph (a nontrivial task for chemical text), (2) matching words against dictionaries to tentatively assign parts of speech, and (3) identifying chemical words, substance identifiers, and units of measurement such as weight, volume, or concentration.

During the lexical phase, the program establishes an internal data structure that facilitates traversing and manipulating the words and phrases of the synthetic description. The data structure allows the assignment of attributes to the words and provides list-manipulation capabilities through a set of subroutines. The data structure consists of list nodes, string nodes, and property nodes. List nodes are connected to each other

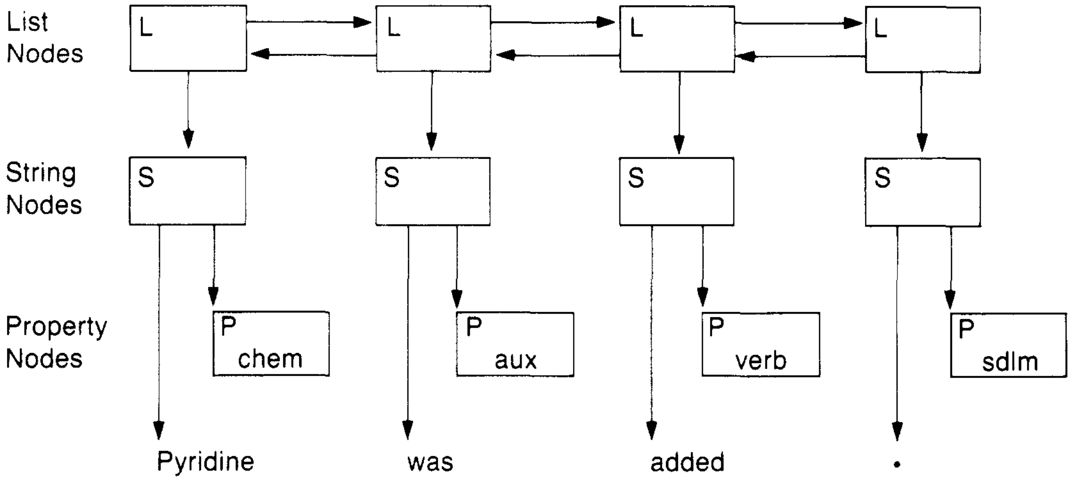

Figure 4. Example of the data structure.

and point to string nodes; string nodes point to a string and its associated property nodes. A string may be a word, a punctuation mark, a numeric character, etc. A property node may be used to associate syntactic or other attributes with a string (for example, the string node for the word "the" can be associated with a property node containing the syntactic code "article"). Property nodes are used to characterize function words, auxiliary verbs, verbs, punctuation, adjectives, adverbs, some nouns, chemical strings, units of measurement, and information to control program execution. In addition, property nodes are used to mark multiword sequences, such as parenthetical expressions and complete chemical substance names. Figure 4 illustrates the internal data structure representation used by the programs.

WORD DEFINITION AND CLASSIFICATION TECHNIQUES

Although a word might simply be defined as a string of alphabetic characters bounded by spaces or other delimiters, this simple definition is not suitable in a chemical context. After analysis of a substantial amount of text, we defined a word as a string of characters between two delimiters. Due to the complexity of the words, it proved easier to define the delimiters than the characters contained within the words. For example, a chemical name may contain embedded commas, parentheses, periods, and hyphens. It is easier to specify the cases when a comma is a delimiter than to describe the format of chemical names that contain commas.

The word delimiters used for our study include the characters space, virgule (slash), semicolon, and percent sign. Other characters must meet specific context constraints before they can be considered delimiters. Thus, the comma, semicolon, and period must be followed by a space. A hyphen must be preceded and followed by at least two alphabetic characters, a right parenthesis must be followed by a space or another punctuation mark, and a left parenthesis must be preceded by a space and followed by an alphabetic or numeric character. These rules make it possible to identify a wide diversity of chemical names without incorrectly segmenting them. The name "5-(2,2-diacetylvinyl)-1,3-dimethyluracil" is considered one word, whereas "heat-treated" is considered to be two words.

For the lexical phase, we created a dictionary of function words and auxiliary verbs to assign syntactic codes to the words of the paragraph. An additional dictionary of common chemical substances, chemical formulas, and chemical word roots supplemented by morphological analysis matches string patterns and applies suffix stemming rules to recognize nouns, adverbs, adjectives, single chemical words, and substance identifiers. The morphological identification of chemical words relies on the identification of name fragments for commonly occurring substances such as "chlor", "ethyl", and "phen", whereas the identification of syntactic categories is based on common endings (e.g., "-ly" for adverbs and "-ness" for nouns). The main function of the lexical phase is to isolate the words of a paragraph describing a chemical reaction and label them with the "most likely" syntactic or descriptive category.

SYNTACTIC PHASE

The syntactic phase of this research is concerned with identifying parts of speech that require context checking and cannot be done by simple dictionary lookup and morphological analysis. The syntactic phase has been implemented by application of "word expert" procedures to the data structures built during the lexical stage. Word experts are procedures whose function is to determine a word's role in a sentence, on the basis of its context.[5]

The structure of text beyond the word level is described by a grammar that specifies the proper construction of prepositional phrases, noun phrases, verb phrases, and higher linguistic structures from the lexical items. The establishment of a grammar is one of the fundamental tasks that must be accomplished before text exhibiting substantial variation, such as natural language text, can be manipulated. The grammar is the basis of the computer programs generated to analyze, or parse, text. In order to represent the syntactic structure of a language in a form that is suitable for a computational task, it is first necessary to formalize the grammar and rid it of any ambiguities.

The grammar employed in our research to describe the synthetic paragraphs of primary journals uses syntax diagrams because they can be visually interpreted more easily than other notations. Use of the word expert approach enables us to implement fragments of a grammar and does not require the development of a comprehensive grammar. This may be illustrated by application of the syntax diagrams in Figure 5 to the following sentence:

A mixture of 1 (1.68 g, 0.01 mol) and acetylacetone
(1.20 g, 0.012 mol) was refluxed for 4 h.

The auxiliary "was" and the verb "refluxed" divide the sentence into two parts designated as noun phrases by the grammar. The noun phrase on the left of the verb consists of the introductory phrase "A mixture" followed by the prepositional phrase with a compound noun "of 1 (...) and acetylacetone (...)". The noun phrase on the right of the verb has no introductory phrase but has the prepositional phrase "for 4 h". Details of the internal constituents of the introductory

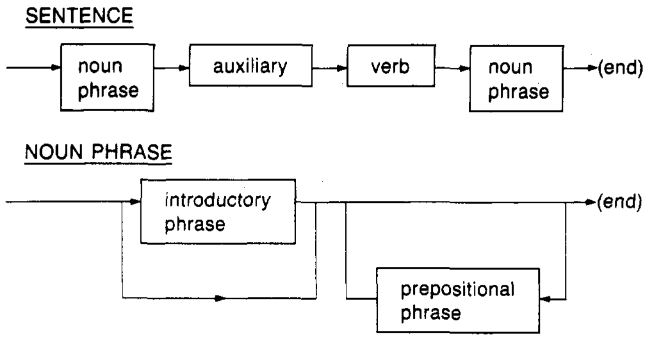

Figure 5. Syntax diagrams for sentence and noun phrases.

and prepositional phrases are not necessary; the phrases are allowed to encompass any word not specified in the formal description. This makes it possible to let the syntactic structure of the sentence dictate the boundaries of chemical substances consisting of several words, some of which may not have been identified by the lexical procedures.

MODEL OVERVIEW

One of the concepts that we have found useful in our research is the development of a model that correlates the text of a synthetic paragraph with a chemical equation. This is useful because chemical equations provide a very natural way for a chemist to describe a chemical reaction. In our simple model, a chemical reaction such as

$$X + Y \rightarrow Z$$

might be expressed as "X and Y were refluxed to give Z", but usually, the form of the sentence is much more complex because quantities and reaction conditions are specified. The discourse structure of the text is superimposed on this model to guide the information extraction process.

Verbs are the focal points used to determine the roles of the chemical substances participating in the reaction. In order to simplify the grammar, complex sentences need to be decomposed into simple sentences containing only one verb. The problem is not only to identify verbs by dictionary lookup but also to distinguish them from verbal clauses, such as participles (past and present) or gerunds. After the sentences are reduced to their simplest form, the context to the left and to the right of the verb may be analyzed by word experts to identify chemical substances by using the syntactic structure of the simple sentence. Prepositions, in coordination with verbs, indicate the chemical role of each chemical substance in a reaction. It is also possible by analysis of the immediate environment of chemical substances in a sentence to identify the quantities and concentration of chemical substances used in a reaction. Reaction conditions, such as time, temperature, and yield, may also be recognized by word expert procedures that operate on clues based on keywords and special punctuation, such as percent and degree symbols.

GENERATION OF A GRAMMAR

The generation of a grammar for a language is an inductive process analogous to what a child must do when learning a language. Models of language acquisition, such as those proposed by Reeker[9] and Langley,[10] involve processes in which rules of a grammar are postulated and modified by experience. In essence, this is the technique used to develop the formal grammar for various constructs in the sentences describing chemical reactions in primary journals.

The basic steps for generating a syntactic description consist of (1) focusing on a feature of the language to be described, (2) collecting examples of text containing the feature, (3) analyzing the text, (4) recognizing patterns, (5) generating a formal description (i.e., syntactic rule) of the pattern, (6) testing with new data to discover exceptions to the pattern, and, finally, (7) repeating the steps until the syntactic rule applies to an adequate portion of the text.

The basic unit of text in the English language is the sentence. However, identification of sentences in text is more complicated than it seems on the surface. Punctuation marks such as periods do not provide reliable markers to delimit sentences because they may also be used to terminate abbreviations or as decimal points in numeric quantities. Another probelm even more fundamental than the identification of sentence boundaries is the decomposition of compound sentences into simple sentences. The main purpose for identifying simple sentences is to simplify the syntax and, consequently, the implementation of the grammar.

Simple sentences can be delimited not only by periods but also by commas, semicolons, or commas followed by conjunctions. According to standard punctuation rules, commas should not be used to separate simple sentences in a compound sentence unless the last of the commas is followed by a conjunction. However, even a comma followed by a conjunction is not a sufficient condition to identify simple sentences since compound subjects can legitimately contain them.

The identification of simple sentences is a complex task. We have defined a simple sentence as a group of words separated by special delimiters and containing one verb (not counting auxiliaries). Once all the words that can be verbs have been marked during the lexical stage, a search is initiated for a sentence delimiter between every two verbs. The search proceeds from right to left to give preference to the rightmost delimiter having the highest rank. Numerical ranks of decreasing value are assigned to semicolon, period, comma followed by conjunction, comma not followed by conjunction, conjunction not preceded by comma, and, finally, the words "which", "that", and "until". Some of these cues will mark dependent clauses in complex sentences; these can be treated like independent clauses and are encompassed by the definition of "simple sentence" above.

The sentence identification procedure changes some of the syntactic roles applied during the lexical phase by correlating the occurrence of words marked as verbs with delimiters such as periods and conjunctions. For example, in a sentence containing "...cooled pyridine was added...", the word "cooled" is recognized as a participle rather than a verb, leaving "added" as the only verb in the sentence (other than the auxiliary).

Following the identification of sentence delimiters, the sentence word expert is applied. Application of the sentence word expert constitutes top-down parsing of a large portion of the grammar for synthetic descriptions because the recognition of chemical substances and the phrases where they occur is described at lower levels of the parse tree. The process is highly efficient due to the identification of verbs and sentence delimiters in the preliminary stage. The sentence word expert describes a large portion of the grammar developed under this project. This word expert focuses on the verb to process the left and right parts of the sentence, introductory and prepositional phrases, and chemical substances and quantities. Figure 5 illustrates the syntax diagram for a sentence in the passive voice and for noun phrases.

The description of the sentence does not differentiate between subject and object and makes it possible to process the noun phrases on either side of the verb with a common subroutine. While prepositions play an important role in the identification of the phrases, parenthetical expressions are ignored during phrase identification.

Chemical substances are identified by parsing the noun phrases into introductory and prepositional phrases. By "introductory phrase" is meant everything preceding the first preposition in the sentence or everything between the verb and

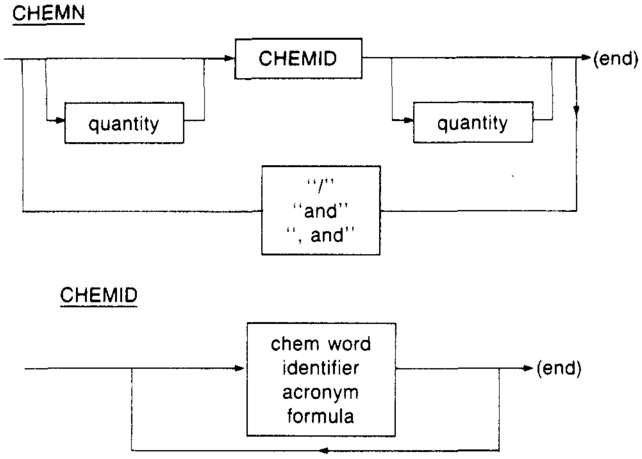

Figure 6. Syntax diagrams for chemical word experts.

the first preposition following it. Each phrase is submitted to analysis by the CHEMN word expert, see Figure 6. Chemical names are complex textual entities consisting of one or more chemical words, acronyms, or identifiers that are contained within introductory or prepositional phrases. Each phrase may contain more than one chemical substance separated by conjunctions and sometimes by slashes, particularly for mixtures of specific ratios.

The CHEMN procedure recognizes complete chemical substances through the CHEMID procedure, which groups adjacent chemical words identified in the lexical phase and uses contextual clues, such as syntactic separators (sentence delimiters, prepositions, etc.) and parenthetical delimiters that specify quantities of the chemical used in the reaction. For JOC and other ACS journals, the names isolated by the program can be verified by reference to the CAS Registry Numbers given at the end of each article.

RECOGNITION OF THE CHEMICAL SUBSTANCE ROLE

Some of the roles that substances may play in a reaction are reactant, product, solvent, etc. These roles are generally implied by the synthetic description and normally are not explicit in the text. This is particularly true for reactants. For example, in the sentence "A solution of A in B was added to C under E to give D.", the substances A and C are reactants, B is a solvent or medium, E is atmosphere, and D is a product. The determination that B is a solvent and A is a solute can be made from the analysis of the context of the word "solution". In this particular case, A is not only a reactant but also it is a solute. Word experts enable us to focus on the context of particular words such as "solution" to identify with a great degree of precision some of the roles of the substances (e.g., solvent, solute). Such word experts take their clues from the prepositional structure (or grammatical case structure) of the phrases comprising the synthetic description. The verb of a sentence provides a focal point that can be advantageously used to determine the components participating in a reaction. Again, the prepositional phrases play an important role in determining the roles of chemical substances because recognition of each substance's role relies substantially on prepositions preceding the substance. For example, the preposition "in" indicates that chemical substances are media, the prepositions "of" or "to" mark chemical substances as reactants, and the preposition "under" followed by the name of a gas indicates that the chemical substance is atmosphere.

The syntax of simple sentences is described in a pragmatic way that facilitates the extraction of chemical substances from the textual description. One of the basic tenets of our approach is that the grammatical framework is the dominant factor mandating the location of chemical components in a sentence. Given a sentence such as "To dry pyridine (100 mL) cooled in a dry ice bath was added methanesulfonyl chloride (34.6 g, 0.302 mol) under nitrogen.", we can illustrate several aspects of substance and role identification. As mentioned above, prepositions provide clues about the roles of the substances. In this example, "pyridine" will be recognized as a reactant, and "dry ice" is recognized as providing information about the temperature. Examination to the right of the verb yields "methanesulfonyl chloride" and "nitrogen" as chemical substances. "Nitrogen" is given the attribute "atomosphere" because it follows the word "under", and "methanesulfonyl chloride" is marked as a reactant consisting of two words and having its associated quantity to its right.

Following application of the CHEMID word expert, the quantity expert scans the environment of the chemical substance for quantity data. Since the quantities can precede or follow a substance, the quantity word expert adds a tag to each substance to indicate whether the quantity occurs to the right or to the left of the substance. Frequently, no quantities are given, particularly for solvents and reagents.

While the quantity word expert is embedded (at a deeper level of the parse tree) within the CHEMN word expert, the temperature and time word experts operate completely independently in a separate pass through the data structure. The time and temperature of a reaction can be isolated by using a fairly simple pattern-matching procedures. Generation of the RIF is the final step that follows the application of all the word expert procedures.

CONCLUSIONS

In the environment that we have created, word experts interact with each other by examining and updating attributes of words or groups of words in the data structure. The application of the word experts is coordinated by a main program that schedules the sequence of their execution and limits the portion of the data structure they must examine. The grammar can handle complex cases, such as those involving multiple substances within the phrases, but it is simplified considerably by the decomposition of complex sentences into simple sentences.

The system has been tested on 40 synthetic paragraphs from the experimental sections of JOC, and 36 were processed satisfactorily. The isolation of chemical substances using the syntactic framework of the sentences achieves 98% recognition of the boundaries of substances. Assignment of the role of the chemical substance in the reaction is currently based on only one grammatical case. Additional grammatical and discourse structure constraints are being implemented, and the results will be published later. Preliminary results indicate, however, that substance role assignments can be achieved automatically for over 80% of the substances.

This work was started in January of 1982, and we have now developed enough word experts to assign syntactic roles to a large percentage of the words and to identify and mark many of the text components that are needed to complete the RIF. Our future efforts will be directed toward developing techniques to increase the reliability of the programs for synthetic paragraphs that follow the simple model and seeking ways of generalizing these techniques to a broader class of text.

REFERENCES AND NOTES

(1) Zamora, E. "Extraction of Chemical Reaction Information using Computational Linguistics Techniques". Symposium on Artificial Intelligence Research and Applications to Chemical Information, 183rd National Meeting of the American Chemical Society, Chemical Information Division, Kansas City, MO, 1982.

(2) Reeker, L. H.; Zamora, E. M.; Blower, P. E. "Specialized Information Extraction: Automatic Chemical Reaction Coding from English Descriptions". "Proceedings of the Conference on Applied Natural

Language Processing", 1983, pp 109-116.
(3) Sager, N. "Natural Language Information Processing"; Addison-Wesley: Reading, MA, 1981.
(4) Rieger, C.; Small, S. "Word Expert Parsing". "Proceedings of the International Joint Conference on Artificial Intelligence 6th", 1979.
(5) Small, S. "Word Expert Parsing". "Proceedings of the Annual Meeting of the Association for Computational Linguistics, 17th", 1979.
(6) Minsky, M. "A Framework for Representing Knowledge". "The Psychology of Computer Vision"; Winston, P., Ed.; McGraw-Hill: New York, 1975.
(7) Cherniak, E. "Organization and Inference in a Framelike System of Common Sense Knowledge". "Theoretical Issues in Natural Language Processing"; Shank, R. C.; Nash-Webber, B. L., Eds.; Mathematical Social Sciences Board: Cambridge, MA, 1975.
(8) Cohen, S. M.; Schermer, C. A.; Garson, L. R. "Experimental Program for On-Line Access to ACS Primary Documents". *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 247-252.
(9) Reeker, L. H. "The Computational Study of Language Acquisition". *Adv. Comput.* **1976**, *15*, 181-239.
(10) Langley, P. "A Model of Early Syntactic Development". "Proceedings of the Annual Meeting of the Association for Computational Linguistics, 20th", 1982.

Extraction of Chemical Reaction Information from Primary Journal Text Using Computational Linguistics Techniques. 2. Semantic Phase[†]

ELENA M. ZAMORA[‡] and PAUL E. BLOWER, JR.*

Chemical Abstracts Service, Columbus, Ohio 43210

Received December 30, 1983

Semantic procedures are described for extracting chemical reaction information from the experimental sections of full papers of American Chemical Society journals. Emphasis is placed on the constraints that the restricted subject domain and limited discourse structure of the synthetic paragraphs have on interpreting the meaning of the words of the text. Frame procedures are used to codify the expectations within the discourse, and case grammar rules determine the role of chemical substances on the basis of these expectations. An experimental system is described that applies case grammar rules and frame procedures to generate Reaction Information Forms, which identify reactants, products, media, and chemical reaction conditions.

INTRODUCTION

The goal of the research described here is to investigate the applicability of computational linguistic techniques to the problem of extracting facts about chemical reactions from the text of primary journals of the American Chemical Society (ACS) and encoding those facts in a form suitable for establishing a reaction database. This research problem was selected as an example of the more general problem of using computational linguistics techniques to create useful databases from files that have accumulated through automated photocomposition procedures. Previous publications have described the lexical and syntactic phases of this work.[1-3] The function of these syntactic phases is to examine the surface characteristics of synthetic paragraphs from the *Journal of Organic Chemistry* (JOC) to obtain as much information as possible about the sentence constituents. The information extracted, however, is limited to the identification of multiword chemical substances from clues from the syntactic structure of the sentences and to the isolation of substance or reaction properties that do not require the use of contextual clues.

The semantic phase of this work tries to deduce the meaning of the synthetic paragraph by identifying the reaction product, reactants, media, and reaction conditions. The model of the text used to describe a chemical reaction determines the way in which a synthetic chemical paragraph is interpreted. A "simple model", based on analysis of a random sample of paragraphs from JOC, is used to identify four components of the paragraph: heading, synthesis, workup, and characterization. The simple model describes the synthesis of a single product and represents the most commonly occurring mode of organization of chemical paragraphs.[2,3] The model makes it possible to use frame procedures to codify the expectations of the discourse structure and of the Reaction Information Form (RIF),[2,3] which is the final representation for the chemical reactions processed by the programs. The frame procedures are used to restrict the scope of case grammar rules, thus improving the assignment of the roles of the substances in the chemical reaction. The following paragraphs discuss briefly some of the terminology that will be used.

BACKGROUND

In many languages, the relationship of a noun to the rest of the sentence is indicated by the different inflections of the noun called "cases". Many distinctions that can be made by inflected forms of words in some languages are made in English by prepositions or word order. However, English still uses declensions for the personal pronouns, which have three cases: nominative, objective, and possessive. In a single sentence with subject, verb, and object such as "I gave it to John.", the nominative case of the pronoun is used to the left of the verb even in passive constructions where subject/object relationships may be altered (e.g., "I was given the book."). The objective case of the pronoun is used to the right side of the verb as in "John gave me the book." or "The book was given to me.". The possessive case has two declensions, one of which has a determiner (adjectival) function and the other a nominal function (e.g., "This is my book." and "This book is mine.").

Although case declensions are used for only a small part of the English language, the term "case grammar" applies to the study of the relationships of noun phrases and verbs as indicated by prepositions that act as "case markers". Case grammars generally associate verbs, which are considered the basic units of the sentence, with their case arguments.[4-6] The arguments are noun phrases or embedded sentences, and they have specific relationships to the verb (e.g., agent, instrument, object); these relationships constitute the cases. The "case structure" of a verb is the set of case relationships that are valid for that verb.

[†] Presented in part before the Division of Chemical Information, 184th National Meeting of the American Chemical Society, Kansas City, MO, Sept 13, 1982.
[‡] Present address: P.O. Box 3554, Gaithersburg, MD 20878.