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.

Frames provide a framework for the organization of knowledge.[7,8] Each frame consists of slots where instantiations of specific expectations can be recorded, and each slot has the characteristics of a subframe that makes it possible to have constrained expectations within a more limited domain. In artificial intelligence, tasks frames have been traditionally applied to simplify the control structure necessary for assigning attributes to conceptual entities. This is generally illustrated in problem-solving situations involving part/whole relationships such that properties of a class recognized at a higher level are inherited by the subframes at lower levels. Thus, if a general description of mammals is available and if it is known that horses are mammals, it is possible to use rules of inference to "understand" many attributes about horses that are typical of mammals from the expectations embodied in the higher level frame.

Although the goals of artificial intelligence are different from those of computational linguistics, there are many common problems in their interfaces with the outside world. The aim of artificial intelligence is to simulate intelligent behavior, for example, in space probes or in robots for industrial environments. Computational linguistics, on the other hand, attempts to understand and generate language with the competence of a human being. The two fields are interrelated because language is a necessary medium for communication for many intelligent tasks and because it is not possible to conceive of understanding language without intelligence.

The correspondence between the cases of a grammar and slots of frames has been noted in recent years,[9-11] and this has made it possible to integrate both approaches, although, traditionally, case grammars have been used for linguistic tasks and frame procedures have been used for artificial intelligence tasks.

## FRAME PROCEDURES

One of the problems that must be confronted in the design of a language-understanding system is how to design the system components and their interaction. Thus, identification of the frames that are to be implemented is a very important consideration. For the extraction of chemical reaction information, the first impulse might be to define a frame containing the desired components: product, reactants, yield, etc. However, consideration of how the substances found in the text will be used to fill the slots of the frame makes it necessary to take into account the discourse structure of primary journal text and the semantic content of the information presented. Thus, the "simple model" integrates the knowledge about the discourse structure with the requirement for the generation of a single synthetic product and provides a logical foundation for the design of the two major frame procedures: the discourse frame and the RIF frame.

Frame procedures provide a set of expectations that have to be fulfilled in particular situations. In analyzing synthetic paragraphs from JOC, the discourse frame embodies the expectation that the synthetic reaction description will contain a heading, synthesis, workup, and characterization. Within each of these sections there are additional expectations for lower order frames; these expectations are that (1) reaction starting materials will be mentioned, (2) reaction products will be mentioned, (3) the efficiency of the reaction will be stated as a yield, and (4) the identity of the product will be characterized by several analytical methods. These frames may contain expectations that are necessary for completeness but may not always be realized, for example, a description of the apparatus, atmosphere, temperature of the reaction, etc.

**Discourse Frame.** The simple model provides a very effective way of assigning roles (e.g., reactant, product) to the substances in synthetic paragraphs because these roles are highly dependent on the context in which the substances occur. However, identification and classification of the sentences of the text into the four categories (heading, synthesis, workup, characterization) requires algorithmic procedures having a detailed knowledge of these four components. This presents the dilemma that the programs to identify the components of the reaction need to be applied in the restricted context for which they are designed, but the identification of the restricted context requires information that is most reliably generated by these same programs.

A bootstrapping procedure breaks this impasse by determining the approximate discourse structure of the synthetic paragraphs by using the list of verbs that are the basis for the case structures in the grammar analysis programs. This look ahead is an effective way of identifying the four components of the discourse structure with minimal computational resources. Care must be taken to identify the discourse components properly to ensure that the expectations embodied in the frame procedures are valid; otherwise, the results will be unreliable.

The basis purpose of the discourse frame is to obtain and make use of locative clues that improve the extraction of information. Knowledge of the context in which specific words occur narrows the scope of their meaning sufficiently to eliminate ambiguities. This may be illustrated by the sentence "3 was added to 2", which interpreted in a general context would most likely refer to an arithmetic operation. In the context of chemical primary journals, however, this represents the combination of two substances identified by numeric referents, which is a common practice. Within this restricted context of chemistry, we can glean further information about the meaning of the verb "add" by using the discourse structure. If this sentence is found in the synthetic description of the discourse, the substances represented by the numbers 3 and 2 are most likely reactants of focal interest to the current reaction; whereas, if the sentence is found in the workup portion of the discourse, it represents a purification step, and no chemical reaction of any importance is implied. The discourse frame slots correspond to the four components of the synthetic paragraph.

**Identification of the Heading.** The heading of a synthetic paragraph generally consists of the name of a chemical substance or of a verbless phrase that may contain one or more chemical substances. The following are examples of typical headings:

reaction of 5c with 2a.

preparation of 4-diazinoindeno(1,2-b)thiophene (5a).

As a rule, the product is given in the heading, but this is not always the case. The discourse analysis identifies the heading by looking for periods (used as delimiters) and then examining the text for verbs. All leading groups of words delimited by periods and containing no verbs are included in the heading.

**Identification of the Synthesis Discourse.** The synthesis discourse is not as easily identified as the heading that precedes it or the text describing the reaction workup that follows it. The heading consists, at most, of a simple phrase, and the characteristic verbs used to describe the reaction workup make that portion of the text easier to identify. We have defined the synthesis discourse as the set of sentences following the heading and preceding the workup, and identify it indirectly by locating the beginning of the workup text. Within the synthesis discourse, we expect to find the starting materials for the reaction, the procedural mechanisms by which the reaction was carried out (including stirring, refluxing, heating, cooling, etc.), the environmental variables that were controlled (such as atmosphere), and the reaction product.

**Table I:** Discourse Frame Expectations

| slot | semantic expectation | syntactic/lexical expectation |
|---|---|---|
| HEADING | product, reactant | no verb |
| SYNTHESIS | product, reactant, media, time, temperature | verbs "add", "reflux", "warm", etc. |
| WORKUP | product, yield | verbs, "crystallize", "wash", "isolate", etc. |
| CHARACTERIZATION | analytical standards, spectral and physical properties | property tags "mp", "bp", "m/e", etc. |

**Identification of the Workup.** The workup is identified by the occurrence of verbs that describe the processes of terminating the reaction and purifying the product for analysis. Some of the verbs that indicate the start of the workup area are crystallized, determined, eluted, extracted, isolated, purified, etc. A list of approximately 60 verbs was sufficient to reliably identify the workup description. Since the workup often involves the addition of chemical substances, for example, to quench the reaction or to dry the reaction mixture, other substances may be found in this body of text, but they are not relevant to the reaction. The end of the workup discourse is marked by the beginning of the characterization section, which is identified by the presence of analytical data.

**Identification of the Characterization.** The characterization contains analytical data that substantiates the synthesis of the reaction product. Physical constants such as melting point, boiling point, and spectral absorption frequencies in the visible, ultraviolet, or infrared ranges occur frequently in the characterization discourse. In addition, proportions of elemental components determined by mass spectral analysis or other analytical techniques are given and compared against the theoretical expectations. Substances that are found in the characterization are expected to be carriers for chromatographic techniques, substances for calibration of analytical instruments, etc. The characterization can be identified on the basis of the abbreviations (mp, IR, UV) of the physical properties measured. Although the text in this part of the discourse is not used in the RIF, it contains a wealth of information that can possibly be used to create numerical databases.

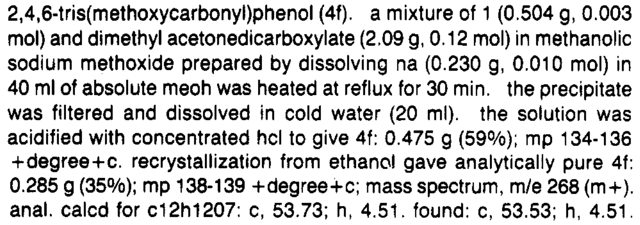
The expectations for the discourse frame can be classified with respect to the semantic components as well as to the syntactic and lexical entities. It is the latter that are used to fill the slots of this frame. The expectations of the discourse frame are summarized in Table I.

**Mapping Procedure for Discourse Frame.** Mapping the text of the synthetic paragraph into the slots of the discourse frame is accomplished by using discourse clues. The following example with paragraph 1 (Figure 1) provides a detailed description of the procedures involved and corresponds to the program implementation.

The heading slot of the discourse frame is filled by scanning for a verb starting from the beginning of the paragraph. The first verb encountered is "heated" since the word "prepared", which can also be a verb, has been identified as a particle during the syntactic stage for this particular example. Once the verb is located, the mapping procedure scans backward, toward the beginning of the paragraph, for a sentence delimiter. When this is found, the portion of text starting from the beginning of the paragraph to the sentence delimiter is placed in the heading slot of the discourse frame:

> HEADING SLOT: 2,4,6-tris(methoxycarbonyl)-phenol (4f).

Filling the synthesis slot of the discourse frame depends upon identifying the verbs used in the workup part of the discourse.



2,4,6-tris(methoxycarbonyl)phenol (4f). a mixture of 1 (0.504 g, 0.003 mol) and dimethyl acetonedicarboxylate (2.09 g, 0.12 mol) in methanolic sodium methoxide prepared by dissolving na (0.230 g, 0.010 mol) in 40 ml of absolute meoh was heated at reflux for 30 min. the precipitate was filtered and dissolved in cold water (20 ml). the solution was acidified with concentrated hcl to give 4f: 0.475 g (59%); mp 134-136 +degree+c. recrystallization from ethanol gave analytically pure 4f: 0.285 g (35%); mp 138-139 +degree+c; mass spectrum, m/e 268 (m+). anal. calcd for c12h1207: c, 53.73; h, 4.51. found: c, 53.53; h, 4.51.

**Figure 1.** Paragraph 1. The data in this and subsequent figures are shown in the form in which they were actually processed. This differs from the form in which they would appear in a journal article in that there is only one type font and some character substitution has occurred; e.g., the degree sign is represented as +degree+.

The mapping procedure scans starting from the end of the heading, looking for any of the various verbs that are characteristic of the workup discourse. When one is found, the procedure scans backward toward the previous sentence delimiter and marks this as the boundary between the synthesis and workup parts of the paragraph. In this example, the verb "filtered" is recognized as a workup verb, and the preceding sentence delimiter becomes the demarcation point to give

> SYNTHESIS SLOT: A mixture of 1 (0.504 g, 0.003 mol) and dimethyl acetonedicarboxylate (2.09 g, 0.12 mol) in methanolic sodium methoxide prepared by dissolving Na (0.230 g, 0.010 mol) in 40 ml of absolute MeOH was heated at reflux for 30 min.

The workup slot of the discourse frame is filled by scanning from the end of the synthesis discourse toward the end of the paragraph, looking for any of the tags used to indicate the physical properties of the product and the analytical techniques used. Some distinctive physical properties typically mentioned in the characterization are melting point, boiling point, mass-to-energy ratio for mass spectra, etc. These properties are identified in the text by abbreviations such as mp, bp, and m/e. The mapping procedure is able to fill out the remaining two slots of the discourse frame by scanning backward from these properties for the preceding sentence delimiter. In this example, the abbreviation for melting point acts as the lexical item that makes it possible to complete the discourse frame slots:

> WORKUP SLOT: The precipitate was filtered and dissolved in cold water (20 mL). The solution was acidified with concentrated HCl to give 4f: 0.475 g (59%);

> CHARACTERIZATION SLOT: mp 134-136 +degree+C. Recrystallization from ethanol gave analytically pure 4f: 0.285 g (35%); mp 138-139 +degree+C; mass spectrum, m/e 268 (m+). Anal. Calcd for C12H1207: C, 53.73; H, 4.51. Found: C, 53.53; H, 4.51.

**Interaction between Discourse Frame and Case Grammar.** The isolation of the four elements of discourse in a synthetic paragraph helps to identify clearly the reactants and products in the paragraph. The preceding example has water, HCl, and ethanol in the workup and characterization slots of the discourse frame. The fact that the substances occur in these parts of the discourse is enough to disqualify them from being identified as reactants or products unless evidence to the contrary is available.

The discourse frame directs the interpretation of the roles of the substances encountered. In the heading, if there are no prepositions, the expectation is that the product only will be mentioned. If prepositions are found, the arguments of the case structures determine the role of the substances, and the default is to treat all substances found as reactants, except where the mapping parameters for the case structures indicate otherwise. Similarly, in the body of the paragraph, substances are assumed to be reactants unless the mapping parameters for the specific verbs directly relate the case arguments to reaction products.

## DEVELOPMENT OF CASE PROCEDURES

Case grammar provides a methodology for analyzing language according to the cases (meanings) associated with each predicate (generally verbs). Prepositions act as case markers that specify the relationships of the case arguments to the predicates. This makes it possible to characterize verbs by case structures that summarize their attributes.

Specific prepositions such as "with" frequently indicate the instrumental case[4,12-14] when used with a verb like "open". Thus, in the sentence "John opened the door with a crowbar", the instrument used to open the door follows the preposition "with". In general, an indefinite number of properties may be specified for a given event. There are important, or significant, properties and there are modifying, or auxiliary, properties, but the difference between them is not always obvious.

Determining the cases that are significant in the context of chemical reactions requires analysis of the properties described by the synthetic paragraph. The description of a chemical reaction generally contains information about the following: reactants; products; media; time; temperature; yield of product; quantities of reactants, media, and products; apparatus; energy; atmosphere.

Reactants, products, and the quantities of these substances must always be present in a description of a chemical reaction and would be considered the main case relations of "chemical reactions". All other relations may be relevant in certain contexts, but they are less commonly specified and are considered to be only modifying relations. Their presence in the chemical reaction description is optional. Notably missing from the list above are categories for reagents and catalysts. At this point, these substances are simply classified as reactants. At a later stage, rules will be added to distinguish between these substance categories.

Another criterion for identifying cases in synthetic descriptions is to divide the descriptions into components with specific discourse characteristics. This makes it possible to take advantage of the fact that the heading of a synthetic paragraph generally contains information about reactants or products; the synthesis discourse contains information about reactants, media, scale, products, time, and temperature; and the workup discourse has information only about yield or scale of product. Thus, the cases can emphasize specific aspects of the structural components of the discourse. Since the implementation of the current system relies on the frame procedures to limit the scope of application of the case grammar, it is necessary to analyze each part of the discourse to derive discourse-specific case rules. The specificity of the rules makes it possible to assign the role of the reaction components with greater precision when the rules are applied within the context for which they were designed. The following paragraphs describe the identification of the cases for the discourse of paragraphs from JOC.

**Identification of Cases for the Heading.** The heading consists of substance names, identifiers, and sentence fragments and requires a different type of analysis than that required for the complete sentences encountered in the synthetic or workup discourse. In the synthetic discourse, the verb is the focal point that determines the identification and assignment of the role of substances in the reaction. The heading, by contrast, may be only a noun phrase but still have many of the complexities of a sentence.

By far the most common form of heading encountered consists of a single substance name followed by its identification number, but titles such as "Conversion of X to Y.", "Chlorination of X.", "Reduction of X to Y.", etc. are not uncommon. In the cases where the title does not contain the name or identification number of the product, the programs have to extract this information from the synthesis or the workup discourse.

In order to try to unify the processing between the heading and the synthetic discourse, headings are considered to consist of noun phrases in the formal sense defined in the syntactic phase of this project.[3] Thus, one introductory phrase followed by zero or more prepositional phrases is expected. The case markers used in the heading are prepositions, and the nouns found in the headings, such as reduction, hydrogenation, etc., serve the same function that verbs serve in the synthetic discourse for the selection of case structures. The case structure for the work "hydrogenation" in the heading coincides to a large degree with the case structure for the verb "hydrogenated" in the synthetic description, but some departures are required in the interpretation of the case markers.

Substances names in the heading, when they appear without any case markers, imply that the substances are products. When case markers are present, the interpretation of the role of the substances is dependent on the case structure imposed by the noun that specifies the reaction process. The main objective for examining the heading is to determine whether it mentions the product, because if not, the product needs to be identified from other parts of the discourse. In sentences such as "Hydrogenation of X.", "Pyrolysis of X.", or "Thermal rearrangement of X.", it is important to know that the product is not in the heading. Similarly, it is essential to recognize that headings, such as "Preparation of X from Y.", and "Conversion of X to Y.", contain the name of the product of the reaction so that the information can be properly recorded. These last examples can be transformed into forms, such as "X was prepared from Y.", that linguistically correspond to the cases for sentences in the synthetic discourse.

**Identification of Case Structures for the Synthesis Discourse.** The semantic analysis tasks pays particular attention to verb arguments generally marked by prepositions that, in coordination with verbal forms, indicate the role of the substances in a reaction. The role of the grammar is to identify the case arguments indicated by the case markers so that they can be mapped to the slots of the RIF frame.

Through manual analysis of 100 synthetic descriptions, it was determined that only 18 common verbs were used to describe a reaction in the synthesis discourse: add, stir, treat, heat, reflux, warm, cool, hydrogenate, suspend, dissolve, bubble, boil, charge, rock, irradiative, keep, prepare, and obtain. The usage of verbs in the paragraphs was studied to determine the case markers associated with them and the type of arguments that verbs take. The verb "add" was the one most commonly encountered, and the following analysis for this verb is representative of the considerations required for most of the others.

The verb "add" is generally used in two ways: (1) To a mixture of A (Q), B (Q), and C (Q) was added a solution of Q of D in Q of X. (2) A solution of A (Q) in Q of X was added to a solution of B (Q) in Q of Y. A, B, C, D, X, and Y are chemical substances, and Q is the quantity of the chemical substances. As illustrated here, the quantity can precede the substance separated by the preposition "of" or it may follow the substance when enclosed in parentheses. In these examples, the subject is not marked by a preposition, but the object is indicated by the preposition "to"; quantities are marked by the preposition "of" or by no preposition at all. Containment is indicated by the preposition "in"; and substances are also marked by the preposition "of". Because the preposition "of" can indicate both a quantity and a chemical substance, the grammar distinguishes these different cases by

**Table II:** Case Structure for "Add"

| predicate | case markers | case arguments |
|---|---|---|
| add | to | object |
| | null | substance |
| | of | substance, quantity |
| | null | quantity |
| | in | containment |

examining the context of the preposition. The role of chemical substances in the reaction is determined by recognizing the associations that the case arguments (objects of the prepositions) have with the semantic expectations expressed by the RIF frame. In both examples for the verb "add", A, B, C, and D are reactants and X and Y are media. These two sentence types have the same case structure for the verb "add". Table II summarizes the components of the case structure for this verb in the synthesis discourse.

We can represent the case structure for the verb "add" as a set of features ⟨S (O) (C) (Q)⟩. This indicates that the subject (S) of the verb "add" must appear in the surface structure of the discourse but the object (O), containment (C), and quanity (Q) are optional. The subject and the object can be instantiated as a chemical substance, or as in a generic descriptor such as mixture, solution, slurry, etc., followed by the preposition "of" and a chemical substance. These generic descriptors always require the preposition "of" to indicate what particular substance is included in the subject or the object.

**Identification of Cases for the Workup.** The workup area of a synthetic paragraph describes the steps for isolating and purifying the product. This part of the discourse expresses the efficiency of the reaction (yield) as a percentage of the amount of product expected theoretically. The yield can be recognized fairly easily from syntactic clues as was described previously.[3] However, the case structure for the workup discourse should also be able to recognize the yield and to associate it with the name of the product. This is important for those cases where the product does not appear in the heading.

Several of the verbs encountered in the workup area that provide clues about the identity of the product are give, afford, yield, and obtained. The verb "give" has the greatest usage and is encountered in active voice and infinitive expressions such as follows: ...to give 4f: 0.475 g (59%); ...to give 184 g (74%) of 5a; ...gave pure 7; ...gave analytically pure 4f: 0.285 g (35%); ...to give X in 64% yield.

The verb "afforded" is used in a very similar way, but the verb "obtained" requires different case markers since the product is to be found in the subject position in passive voice constructions.

## CONSOLIDATION OF CASE STRUCTURES

A fact that becomes evident after examining the case structures for several verbs is that the prepositions used as case markers for one verb are used for other verbs in the same way and correspond to the same case arguments. For example, the verbs "add" and "stir" both use the preposition "of" as an indicator of a substance name. Although the case structures for both verbs differ because the verb "stir" has more case markers, there are no conflicts in the ways that case markers correspond to the case arguments.

As long as there are no conflicts, it is possible to create a combined case structure representative of the combined verbs. The resulting case structure however, will have more case markers than can possibly occur for a single verb but has advantage of being simpler and less redunant to implement than several case structures with minor differences.

The verbs of the synthetic paragraphs for JOC were consolidated into two distinct case structures. The verbs add, stir, treat, heat, reflux, warm, cool, irradiation, rock, keep, hydrogenate, suspend, dissolve, bubble, boil, and charge were combined into one case structure, and the verbs prepare and obtain were combined into a second one. The major differences between the combined case structure for the verbs "prepare" and "obtain" from the case structure for the other verbs is indicated by asterisks in Table III. The mapping parameter is used exclusively to associate the subject of "prepare" and "obtain", and for that of the semantic expectation "product" as will be described later. Similarly, the case marker "from" is associated with these two verbs only.

**Table III:** Combined Case Structure

| case markers | case arguments | mapping parameter* |
|---|---|---|
| null | subject | product* |
| to | object, temperature | |
| null | quantity | |
| of | substance | |
| in | containment | |
| at | temperature, pressure | |
| for | time | |
| under | atmosphere | |
| with | object | |
| into | object | |
| through | object | |
| by | process | |
| from* | object | |

**Table IV:** Frame for the Reaction Information Form

| | |
|---|---|
| product slot: | product quantity slot: |
| reactant slots: | reactant quantity slots: |
| media slots: | media quantity slots: |
| time slot: | |
| temperature slot: | |
| yield slot: | |

Although the combination of the case markers into a single case structure creates some apparent ambiguities, the context makes it possible to be very precise. As an example, consider the case marker "to" and its case arguments, which can be an object or a temperature. Since temperatures following the preposition "to" are numeric quantities followed by a degree sign, it is virtually impossible to confuse them with a substance name or any other object of the preposition. The same is true for the preposition "at", which can be used to indicate the pressure used during the reaction or a temperature.

## REACTION INFORMATION FORM FRAME

The RIF frame, unlike the discourse frame, has a fixed number of categories and a variable number of slots. The categories of this frame correspond to the (1) products, (2) reactants, (3) media, (4) time, (5) temperature, and (6) yield. The slots of the RIF frame correspond to each of the above categories but permit one or more instances of each category to occur. This is important since an unspecified number of reactants, products, or solvents may be involved in a reaction. Although the RIF frame permits more than one product, the simple model restricts the product slot to only one entry. In addition, associated with each of the slots for the products, reactants, and media are slots for recording the quantity of each substance. Table IV illustrates the slots of the RIF frame.

The RIF frame expresses our expectations of the information that should be present in a synthetic description paragraph from JOC. In contrast to the discourse frame, which directs the interpretation of the text by limiting the application of the case structures to specific subunits of the synthetic paragraph, the RIF frame serves to coordinate the data identified by the case structure procedures into a coherent RIF by the use of mapping procedures.

The RIF frame is applied after all the case structure procedures have executed and identified the text components. The slots of the RIF frame indicate the expectation that associated with every substance there will be a quantity. Thus, as the data structure is scanned for chemical substances, the corresponding quantities are also extracted to fill the appropriate slots. The association of the quantities with the substances is made by the case structure procedures that add the tags "lqty" and "rqty" to each substance to indicate whether the associated quantity occurs to the left or to the right of the substance, respectively, as illustrated here:

```
A solution of quinine (0.4 g) in 40 ml of methanol ...

              ======= -------    -----    ========

              rqty --> qty       qty  <-- lqty
```

**Mapping Procedure for RIF Frame.** The function of the RIF mapping procedures is to associate the arguments of the case structures for each specific verb or noun with the corresponding semantic entities needed to fill the RIF frame slots. Since the mapping is specific for each case structure, it is advantageous to associate an RIF mapping parameter with each argument of a case structure. This may be best illustrated by an example for the verb "obtain".

In a sentence such as "P (Qp) was obtained by adding Qb of B to Qc of C.", the case structure for "obtain" generates the following syntactic associations:

| predicate | case markers | case arguments | instance |
|---|---|---|---|
| obtain | null | subject | P (Qp) |
| | by | process | adding |
| | null | object | Qb of B |
| | to | object | Qc of C |

The RIF mapping parameter, as illustrated in the combined case structure given earlier, is an additional column that equates "subject" with "product" and "object" with "reactant" (the latter by default). The mapping varies from verb to verb, so that for the verbs "add", "boil", and "stir" both "subject" and "object" would be mapped to "reactants", while the mapping parameters for the verb "produced" would be similar to those for the verb "obtain".

The fact that the synthetic paragraphs of JOC represent a restricted domain of discourse makes it possible to incorporate the mapping procedures within the framework of the case structures. In a less restrictied domain, this would not be possible. Furthermore, the restricted domain of synthetic chemistry allows us to develop mapping procedures that would be problematic in a more general setting. The word "solution", for example, is not likely to mean "resolve a problem" in our limited domain. Invariably, the word "solution" refers to the physical act of dissolving a substance in a solvent. Consequently, mapping procedures can be developed to identify the solutes and the solvents from the case structures and then, by implication, to assign the solutes to the reactant slots and the solvents to the media slots of the RIF frame.

The mapping procedure for the RIF frame also fills the slots for the time, temperature, and yield. These three constituents of the RIF are recognized by using syntactic criteria and then labeled in the data structure. Since multiple temperature conditions can be specified in the paragraph, the mapping procedure collects all the temperatures tagged within the synthetic discourse and places them in the RIF frame slot.

## ANALYSIS OF PROGRAM RESULTS

In this section, several examples will be examined to illustrate the program performance and to point out some of the problems that still remain to be solved. This analysis of the results will examine program failures to try to establish (1) deficiencies that can be corrected by ordinary programming techniques and (2) problems that require more sophisticated computational linguistics techniques.

**Table V:** Reaction Information Form for Paragraph 1

| role | amount | substance name |
|---|---|---|
| product | | 2,4,6-tris-(methoxycarbonyl)phenol |
| reactant | (0.504 g, 0.003 mol) | 1 |
| reactant | (2.09 g, 0.12 mol) | dimethyl acetonedicarboxylate |
| medium | | methanolic sodium methoxide |
| medium | 40 ml | meoh |
| time | 30 min | |
| temp | heated | reflux |
| yield | 59% | |

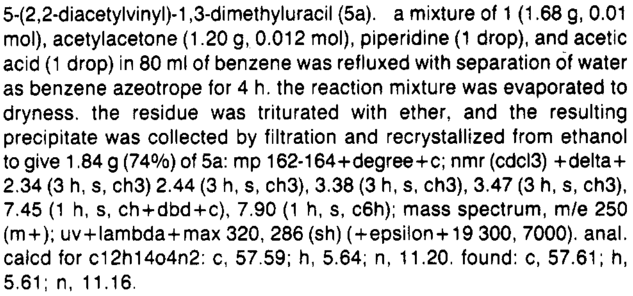
The end of this section presents performance statistics of the current programs. The diversity of expression possible in natural language is so vast that no program can successfully handle 100% of all cases. Yet, it is possible to achieve results that can eliminate much of the tedious work of manually identifying and keyboarding information to build a database.

The RIF generated for paragraph 1, see Table V, obtained the reaction product from the heading. The reactants, numerical identifier 1, dimethyl acetonedicarboxylate, and their corresponding quantities were isolated from the first sentence. Methanolic sodium methoxide and methanol were identified as the medium because of the occurrence of the preposition "in". The time and temperature are obtained from the first sentence, and the yield is identified from the third sentence of the paragraph.

Overall, the information on this RIF is filled out correctly. The designation of methanolic sodium methoxide as a medium is not strictly correct, but in a manually aided database-building environment, this entry could easily be eliminated. The treatment of sodium methoxide in this case also brings up some interesting problems. The preposition "by" in the phrase "by dissolving na...(in)...methanol" indicates that the sodium methoxide is prepared in situ by some process. A chemist can further recognize that it is a reagent prepared from sodium and methanol and that methanol is both the medium and one source of the reagent. We can compare this with another common situation, that of preparing a Grignard reagent in situ from Mg and, say, methyl bromide. In this case, both substances are simply sources of the Grignard reagent, and the RIF should include the active reagent, or its sources, but not all three substances. These two contrasting examples illustrate some of the difficulties in analyzing text describing reagents or reactants prepared in situ. The substances that should be included on the RIF and their reaction role assignments cannot be determined simply from the structure of the textural description but require a deeper analysis of the chemistry involved.

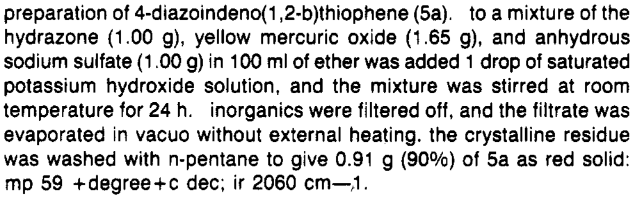
By their very nature, the synthesis paragraphs that this program is designed to analyze provide a large number of clues to their semantic content. The chemical nomenclature, numerical data, and unit abbreviations are the most useful. But the considerations in the foregoing paragraph indicate the depth of analysis required to resolve some of the problems encountered in filling out an RIF. A simple matching of the surface patterns is not adequate. A sophisticated and complex mechanism, involving both linguistic analysis and a knowledge of chemical principles, will be needed.

The RIF generated from paragraph 2 (Figure 2) incorporates elements from various parts of the discourse, see Table VI. The product name is obtained from the heading. All the substances and quantities identified as "reactants" were obtained from the first sentence. Benzene and water, identified as media, and the time and temperature were also derived from the first sentence. The yield was obtained from the third sentence.



5-(2,2-diacetylvinyl)-1,3-dimethyluracil (5a). a mixture of 1 (1.68 g, 0.01 mol), acetylacetone (1.20 g, 0.012 mol), piperidine (1 drop), and acetic acid (1 drop) in 80 ml of benzene was refluxed with separation of water as benzene azeotrope for 4 h. the reaction mixture was evaporated to dryness. the residue was triturated with ether, and the resulting precipitate was collected by filtration and recrystallized from ethanol to give 1.84 g (74%) of 5a: mp 162-164+degree+c; nmr (cdcl3) +delta+ 2.34 (3 h, s, ch3) 2.44 (3 h, s, ch3), 3.38 (3 h, s, ch3), 3.47 (3 h, s, ch3), 7.45 (1 h, s, ch+dbd+c), 7.90 (1 h, s, c6h); mass spectrum, m/e 250 (m+); uv+lambda+max 320, 286 (sh) (+epsilon+19 300, 7000). anal. calcd for c12h14o4n2: c, 57.59; h, 5.64; n, 11.20. found: c, 57.61; h, 5.61; n, 11.16.

**Figure 2.** Paragraph 2.



preparation of 4-diazoindeno(1,2-b)thiophene (5a). to a mixture of the hydrazone (1.00 g), yellow mercuric oxide (1.65 g), and anhydrous sodium sulfate (1.00 g) in 100 ml of ether was added 1 drop of saturated potassium hydroxide solution, and the mixture was stirred at room temperature for 24 h. inorganics were filtered off, and the filtrate was evaporated in vacuo without external heating. the crystalline residue was washed with n-pentane to give 0.91 g (90%) of 5a as red solid: mp 59 +degree+c dec; ir 2060 cm—,1.

**Figure 3.** Paragraph 3.

**Table VI:** Reaction Information Form for Paragraph 2

| role | amount | substance name |
|---|---|---|
| product | | 5-(2,2-diacetylvinyl)-1,3-dimethyluracil |
| reactant | (1.68 g, 0.01 mol) | 1 |
| reactant | (1.20 g, 0.0012 mol) | acetylacetone |
| reactant | (1 drop) | piperidine |
| reactant | (1 drop) | acetic acid |
| medium | 80 ml | benzene |
| medium | | water |
| time | 4 h | |
| temp | refluxed | |
| yield | 74% | |

**Table VII:** Reaction Information Form for Paragraph 3

| role | amount | substance name |
|---|---|---|
| product | | 4-diazoindeno(1,2-b)thiophene |
| reactant | (1.00 g) | hydrazone |
| reactant | (1.65 g) | mercuric oxide |
| reactant | (1.00 g) | sodium sulfate |
| medium | 100 ml | ether |
| reactant | 1 drop | potassium hydroxide |
| time | 24 h | |
| temp | room temperature | |
| yield | 90% | |

**Table VIII:** Summary of Text Results

| RIF field name | no. of occurrences | % identified correctly |
|---|---|---|
| product | 81 | 84 |
| reactant | 211 | 95 |
| media | 105 | 95 |
| atmosphere | 51 | 100 |
| quantities | 243 | 93 |
| temperature | 147 | 98 |
| time | 88 | 93 |
| yield | 81 | 96 |

Although "water" is identified as a medium, a chemist would consider it an undesirable reaction byproduct or contaminant in this case. This can be deduced from the fact that water is removed ("separated") as a benzene azeotrope.

The identification of piperidine and acetic acid as reactants is also questionable in this RIF. It is unusual that these two chemical substances are measured in "drops". Without additional knowledge of chemistry, it can be deduced that these substances are more likely to be stabilizers or catalysts for the reaction rather than reactants, and the use of informal units (drops) reinforces this conclusion. But the precise determination of the role of these substances would require analysis of the reaction at a structural level.

A question that must be asked is whether the RIF would be useful in its present form. The answer has to be affirmative, since there is no misleading information. The fact that both benzene and water are listed as media would be acceptable in a computerized search of such a database. The fact that piperidine and acetic acid are listed as reactants would enable this reaction to be retrieved as a possible hit when these substances are specified as reactants. The determination of whether this is indeed a hit or a false drop would depend on the scope of the query, rather than on the information recorded for the RIF.

Except for the yield, the RIF for paragraph 3 (Figure 3) was derived from the heading and from the first sentence, see Table VII. In this example, the most notable problem is that "hydrazone", which is listed as a reactant, is not a reactant at all. The problem occurs because the paragraph describes one of the starting materials as "the hydrazone", which is a specific hydrazone described earlier in the publication and which is referenced in this paragraph by its generic structural characteristics. The fact that the author has not used a numeric identifier makes it possible to solve the problem easily. Instead, it would be necessary to scan the preceding portion of the paper for a hydrazone that would be a plausible starting material for this reaction.

The mechanism required for inferring that "hydrazone" is not a plausible reactant in this paragraph is very complex. As a general rule, articles such as "the" are not used to refer to specific chemicals. This can provide a triggering mechanism to indicate that there is something unusual. How to proceed from this point will require additional research.

It is important to question whether something has been lost by not including the word "yellow" as part of the name of the second reactant. Would it have been better to list "yellow mercuric oxide"? Does the word "yellow" refer only to the color of the mercuric oxide or does it imply a different chemical composition in the same way that "white lead" indicates the compound lead carbonate rather than lead that is white in color? This question is beyond the scope of the program, but it can be easily solved by increasing the lexicon. The program currently drops the word yellow because it is an adjective and contains no chemical word roots. The same would be true if the paragraph had described "hot mercuric oxide".

The omission of the word "anhydrous" from "sodium sulfate" could be subjected to the same arguments as those for the word "yellow", but in this case a decision was made when the lexicon was built to exclude the concept of "dry" or anhydrous from chemical names except when it is part of the substance name, such as "dry ice".

The results of tests on 81 paragraphs describing synthetic reactions that followed the simple model are summarized in Table VIII. This table describes the number of occurrences of each field of the RIF extracted automatically and gives a percentage of the success of the process.

The number of occurrences for many fields is greater than the number of paragraphs processed because, usually, at least two reactants are used in each reaction description. Quantities are given for every reactant, product, and medium. Many reaction descriptions also specify several time and temperature conditions. The percentage of products identified correctly can be improved by modifying the procedures invoked when there is no heading in the paragraph or the product is not given in the heading. This is the subject of further work.

## CONCLUSIONS

The object of the research reported here was to evalute the application of computational linguistics techniques for the extraction of chemical reaction data from primary journal text to create useful databases. The results obtained from this project provide data that will eventually be useful in comparing the value and the cost of the information obtained in this way with that available through full-text searching. It is necessary to evaluate whether the cost of processing the data is justified on the basis of the value added by the high precision of the results and whether some of this information might also be retrievable through text searching of the primary journal text.

The techniques explored in the semantic phase of this project include the use of a case grammar and the use of frames to map the surface structure of the text into an internal representtion from which an RIF can be printed. It would have been possible to use other techniques, but it is not evident that they could have given better results. The "frames" could have been replaced with "scripts"[15,16] without any significant differences in methodology. If a case grammar had not been used, the most fruitful alternatives might have used either a phrase structure grammar or montauge grammer with very similar implementations.

What are the prospects for automatically creating a chemical reaction database? It is not deemed possible that such a database could be created completely automatically with the techniques developed under this project because the total number of errors would be too high to obtain a useful scientific database. However, the number of errors per reaction description is low enough that it is possible to manually edit the data generated by the program with a fraction of the human resources required to generate the complete database. From this point of view, the techniques explored here can have practical application in the not too distant future.

Before this technique can be applied on a large scale to ACS primary journals, it will be necessary to investigate more sophisticated models of the synthetic paragraphs. Since complex paragraphs may contain more than one product, substantial analysis will be required to handle such paragraphs satisfactorily. In addition, some mechanical procedures such as the retrieval of the CAS Registry Number and chemical name from the tables included at the end of the journal articles will need to be implemented.

Some problems inherent in the current approach cannot be corrected easily. The most notable problem is the fact that the system has little knowledge about the meaning of a chemical reaction. For a system that attempts to extract chemical reaction information, this is a severe deficiency. As discussed earlier, textual clues do not provide sufficient information for the system to be able to discern consistently whether a substance is a catalyst, a byproduct, or a reactant. By incorporating a model of a chemical reaction in the program (or by applying it to the RIF generated), many inconsistencies could be resolved. This would involve adding to the program knowledge about stoichiometry, mass balance, and chemical structure rearrangements, the uses of common chemicals in the laboratory, and some general concepts of chemistry.

In summary, the techniques studied here can be regarded with cautious optimism as a mechanism for generating databases from primary journal text. Much additional work remains to be done to refine the techniques so that they work with the reliability required to build a commercial database, but the use of the programs as a mechanical aid to reduce the cost of manual database building will probably be the first practical application.

## REFERENCES AND NOTES

(1) Reeker, L.H.; Zamora, E. M.; Blower, P. E. "Specialized Information Extraction: Automatic Chemical Reaction Coding from English Descriptions". Conference on Applied Natural Language Processing, Santa Monica, CA, Feb 1–3, 1983.

(2) Zamora, E. M. "Extraction of Chemical Reaction Information using Computational Linguistics Techniques". Symposium on Artifical Intelligence Research and Applications to Chemical Information, 184th National Meeting of the American Chemical Society, Chemical Information Division, Kansas City, MO, Sept 13, 1982.

(3) Zamora, E. M.; Blower, P. E., Jr. "Extraction of Chemical Reaction Information from Primary Journal Text Using Computational Linguistics Techniques. 1. Lexical and Syntactic Phases". *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 176–181.

(4) Bruce, B. "Case Systems for Natural Language". *Artif. Intell.* **1975**, *6*, 327–360.

(5) Fillmore, C. J. "The Case for Case". "Universals in Linguistic Theory"; Bach, E.; Harms, R., Eds.; Holt, Rinehart, and Winston: New York, 1968.

(6) Fillmore, C. J. "Types of Lexical Information". "Semantics: An Interdisciplinary Reader"; Stainberg, D. D.; Jakobovits, L. A., Eds.; Cambridge University Press: London, 1971.

(7) Minsky, M. "A Framework for Representing Knowledge". "The Psychology of Computer Vision"; Winston, P., Ed.; McGraw-Hill: New York, 1975.

(8) Nayvelt, E. M. "Method of Frames for Automatic Recognition of Chemical Text". Thesis for the Conference on Automatic Processing of Text, Makhachkala, USSR, July 1978.

(9) Charniak, E. "The Case-Slot Identity Theory". *Cognit. Sci.* **1981**, *5*, 285–292.

(10) Fillmore, C. J. "The Case for Case Reopened". "Syntax and Semantics 8: Grammatical Relations"; Cole, P.; Sadock, J. M., Eds.; Academic Press: New York, 1977.

(11) Winston, P. H. "Learning by Creating and Justifying Transfer Frames". "Artificial Intelligence: An MIT Perspective"; Winston, P. H.; Brown, R. H., Eds.; MIT Press: Cambridge, MA, 1979; Vol. 1, pp 345–374.

(12) Grimes, J. "The Thread of Discourse". "NSF Technical Report 1"; Cornell University: Ithaca, NY, 1972.

(13) Sidner, C. "Disambiguating References and Interpreting Sentence Purpose in Discourse". "Artificial Intelligence: An MIT Perspective"; Winston, P. H.; Brown, R. H., Eds.; MIT Press: Cambridge, MA, 1979; Vol. 1, pp 231–252.

(14) Simmons, R. F. "Semantic Networks: Their Computation and Use for Understanding English Sentences". "CAI Lab Report NL-6"; Computer Science Department, The University of Texas: Austin, TX, 1972.

(15) Schank, R. C.; Ableson, R. P. "Scripts, Plans, Goals, and Understanding"; Lawrence Erlbaum: Hillsdale, NJ, 1977.

(16) Schank, R. C. "Language and Memory". *Cognit. Sci.* **1980**, *4*, 243–284.