SPECIALIZED INFORMATION EXTRACTION: AUTOMATIC CHEMICAL REACTION CODING FROM ENGLISH DESCRIPTIONS

Larry H. Reeker* Elena M. Zamora** and Paul E. Blower

Chemical Abstracts Service
2540 Olentangy River Road
P.O. Box 3012
Columbus, Ohio 43210

ABSTRACT

In an age of increased attention to the problems of database organization, retrieval problems and query languages, one of the major economic problems of many potential databases remains the entry of the original information into the database. Specialized information extraction (SIE) systems are therefore of potential importance in the entry of information that is already available in certain restricted types of natural language text. This paper contains a discussion of the problems of engineering such systems and a description of a particular SIE system, designed to extract information regarding chemical reactions from experimental sections of papers in the chemical literature and to produce a data structure containing the relevant information.

I. INTRODUCTION

A. Overview of the Paper

In an age of increased attention to the problems of database organization retrieval problems and query languages, one of the major economic problems of many potential databases remains the entry of the original information into the database. A large amount of such information is currently available in natural language text, and some of that text is of a highly stylized nature, with a restricted semantic domain. It is the task of specialized information extraction (SIE) systems to obtain information automatically from such texts and place it in the database. As with any system, it is desirable to minimize errors and human intervention, but a total absence of either is not necessary for the system to be economically viable.

In this paper, we will first discuss some general characteristics of SIE systems, then describe the development of an experimental system to assist in the construction of a database of chemical reaction information. Many journals, such as the *Journal of Organic Chemistry*, have separate experimental sections, in which the procedures for preparing chemical compounds are described. It is desired to extract certain information about these reactions and place it in the database. A reaction information form (RIF) was developed in another project to contain the desired information. The purpose of the system is to eliminate the necessity in a majority of cases, for a trained reader to read the text and enter the RIF information into the machine.

B. Some Terminology

In the discussion below, we shall use the term *grammar* to mean a system consisting of a *lexicon*, a *syntax*, a *meaning representation* language, and a *semantic mapping*. The lexicon consists of the list of words in the language and one or more grammatical categories for each word. The syntax specifies the structure of sentences in the language in terms of the grammatical categories. Morphological procedures may specify a "syntax" within classes of words and thereby reduce the size of the lexicon. A discourse structure, or extrasentential syntax, may also be included.

The semantic mapping provides for each syntactically correct sentence a meaning representation in the meaning representation language, and it is the crux of the whole system. If the semantic mapping is fundamentally straightforward, then the syntactic processing can often be reduced, as well. This is one of the virtues of SIE systems: Because of the specialized subject matter, one can simplify syntactic processing through the use of *ad hoc* procedures (either algorithmic or heuristic). In many cases, the knowledge that allows this is nonlinguistic knowledge, which may be encoded in *frames*. Although this is not always the sense in which "frame" is used, this is the sense in which we shall use the term in our discussion below: Frames encode nonlinguistic "expectations" brought to bear on the task. In this light, it is interesting to explore the subject of case-slot identity, as raised by Charniak (1981). If the slots

* Current address: Department of Computer Science, Tulane University, New Orleans, Louisiana 70118.

** Current address: P.O. Box 3554, Gaithersburg, Maryland 20278.

are components of frames, and cases are names for arguments of a predicate, then the slots in any practical language understanding system may not correspond exactly to the cases in a language. In fact, the predicates may not correspond to the frames. On the other hand, if the language is capable of expressing all of the distinctions that can be understood in terms of the frames, one would expect them to grow closer and closer as the system became less specialized. The decision as to whether to maintain the distinction between predicate/case and frame/slot has a "Whorfian" flavor to it. We have chosen to maintain that distinction.

Despite the general decision with regards to predicates and slots, some of the grammatical categories in our work do not correspond precisely to conventional grammatical categories, but are specialized for the reaction information project. An example is "chemical name". This illustrates another reason that SIE systems are more practical than more general language understanding systems: One can use certain *ad hoc* categories based upon the characteristics of the problem (and of the underlying meanings represented). This idea was advocated several years ago by Thompson (1966) and used in the design of a specialized database query system (DEACON). Its problem in more general language processing applications - that the categories may not extend readily from one domain to another and may actually complicate the general grammar - does not cause as much difficulty in the SIE case. The danger of using *ad hoc* categories is, of course, that one can lose extensibility, and must make careful decisions in advance as to how specialized the SIE system is going to be.

II. SPECIALIZED INFORMATION EXTRACTION

A. Characteristics of the SIE Task

The term "specialized information extraction" is necessarily a relative one. Information extraction can range from the simplest sorts of tasks like obtaining all names of people mentioned in newspaper articles, to a full understanding of relatively free text. The simplest of these require of the program little linguistic or empirical knowledge, while the most complex require more knowledge than we know how to give. But when we refer to an SIE task, we will mean one that:

(1) Deals with a restricted subject matter

(2) Requires information that can be classified under a limited number of discrete parameters, and

(3) Deals with language of a specialized type, usually narrative reports.

SIE programs are more feasible than automatic translation because the restrictions lessen the ambiguity problems. This is even true in comparison to other tasks with a restricted subject matter, such as natural language computer programming or database query. Furthermore, these latter tasks require a very low error rate in order to be useful, because users will not tolerate either incorrect results or constant queries and requests for rewording from the program, while SIE programs would be successful if they produced results in, say, 80% of cases and required that the information extraction be done by humans in the others. Even small rates of undetected errors would be tolerable in many situations, though one would wish to minimize them

The lessened syntactic variety in SIE tasks means that the amount of syntactic analysis needed is lessened, and also the complexity of the machinery for the semantic mapping. At the same time, the specialized semantic domain allows the use of empirical knowledge to increase the efficiency and effectiveness of analysis procedures (the lessening of ambiguity being only one aspect of this).

The particular cases of SIE that we have chosen are highly structured paragraphs, describing laboratory procedures for synthesizing organic substances which were taken from the experimental section of articles in *J. Org. Chem.* Our feeling is that the full text of chemical articles is beyond the state of the SIE art, if one wants to extract anything more than trivial information; but the limited universe of discourse of the experimental paragraphs renders SIE on them feasible.

B. The Engineering of SIE Systems

Since the days of the early mechanical translation efforts, the amount of study of natural language phenomena, both from the point of view of pure theory and of determining specific facts about languages, has been substantial. Similarly, techniques for dealing with languages and other sorts of complex information by computer have been considerably extended and the work has been facilitated by the provision of higher-level programming languages and by the availability of faster machines and increased storage. Nevertheless, the state of scientific knowledge of language and of processes for utilizing that knowledge is still such that it is necessary to take an "engineering approach" to the design of computational linguistics systems.

In using the term "engineering", we mean to indicate that compromises have to be made in the design of the system between what is theoretically desirable, and what is feasible at the state of the art. Failing to have a complete grammar of the language over which one wishes to have SIE, one uses heuristics to determine features that one wants. At the same time, one uses the scientific knowledge available, insofar as that is feasible. One builds and tests model or pilot systems to explore problems and techniques and tries to extrapolate the experience to production systems, which themselves are likely to have to be "incrementally developed".

In any engineering context, evaluation measures are important. These measures allow one to set criteria for acceptability of designs which are likely always to be imperfect, and to compare alternative systems. The ultimate evaluation measure on which management decisions rest is usually cost/benefit ratio. This can be determined only after examining the human alternatives and their effectiveness. It is important to be able to quantify these alternatives, and this is often not done. For instance, it is common to assume that an automatic system should not produce errors, whereas humans always do; so the percentage of errors should be determined experimentally in each case and compared.

For the evaluation of SIE systems, we would like to propose three measures:

(1) Robustness - the percentage of inputs handled. Most real SIE systems will reject certain inputs, so the robustness will be one minus the percentage rejected.

(2) Accuracy - the percentage of those inputs handled which are correctly handled.

(3) Error rate - the percentage of erroneous entries within incorrectly an handled input.

Probably the most difficult aspect of SIE engineering is the provision of a safety factor - an ability of the system to recognize inputs that it cannot handle. It is clear that one can create a system that is robust and acceptably accurate which has unacceptable error rates for certain inputs. If the system is to be useful, it must be possible automatically to determine which documents contain unacceptable error rates. It does no good to determine this manually, since that would mean essentially redoing all of the information extraction manually, and the space of documents is not sufficiently uniform or continuous that sampling methods would do any good. It appears, then, that the only way that one is going to be able to provide a safety factor is to have a system that understands enough about the linguistic and nonlinguistic aspects of the texts to know when it is not understanding (at least most of the time). We shall have more to say about the safety factor when we discuss our system below.

One suggestion often made for "intelligent" systems is that they be given some provision for improving their performance by "learning". Generally the problem with this suggestion is that the complexity of the learning process is greater than that of the original system, and it is also unclear in many cases what the machine needs to learn. It nevertheless seems feasible for SIE systems to learn by interaction with people who are doing information extraction tasks. The simplest case of this would be augmenting the lexicon, but others should be possible. The first step in this process would be to build in a sufficient safety factor that most incorrectly handled documents can be explicitly rejected. The second would be to localize the factors that caused the rejection sufficiently to be able to ask for help from the person doing the manual extraction process. Although we have considered this aspect of SIE development, we have not made any attempt to implement it.

A. The Description of Chemical Reactions

A particular task that would appear to be a candidate for SIE, under the criteria given above, is the extraction of information on chemical reactions from experimental sections of chemical journals. The journal chosen for our experimental work was the Journal of Organic Chemistry. Two examples of reaction descriptions from this journal are shown in Figure 1. Both of these examples have a particular type of discourse structure, which we have called the "simple model". The paragraphs in the figure (but not in the actual texts) are divided into four components: a heading, a synthesis, a work-up, and a characterization. Usually, the heading names the substance that is produced in the reaction, the synthesis portion describes the steps followed in conducting the reaction, the work-up portion describes the recovery of the substance from the reaction mixture, and the characterization portion presents analytical data supporting the structure assignment. Most of the information that we wish to obtain is in the synthesis portion, which describes the chemical reactants, reaction conditions and apparatus.

Figure 2 shows the Reaction Information Form (RIF) designed to hold the required reaction information, with information supplied for the two paragraphs illustrated in Figure 1. One point to notice is that not every piece of data is contained in every reaction description. Thus there are blanks in both examples, corresponding to information left unspecified in the corresponding reaction descriptions (those shown in Figure 1).

B. An SIE System for Reaction Information

1. General Organization

The chemical reaction SIE is written in PL/I and runs on a 370/168 under TSO. The testing of certain of the algorithms and heuristics has been done using SNOBOL4 (SPITBOL) running under UNIX on a PDP 11/70. The choice of PL/I on the 370 was dictated by practical considerations involving the availability of textual material, the unusual format of that material, and the availability of existing PL/I routines to deal with that format.

The programs comprising each stage of the system are implemented modularly. Thus the lexical stage involves separate passes for individual lexical categories. In some cases, these are not order-independent. In the syntactic phase, the individual modules are "word experts", and in the last (extraction) phase, they are individual "frames" or components of frames.

2. The Lexical Stage

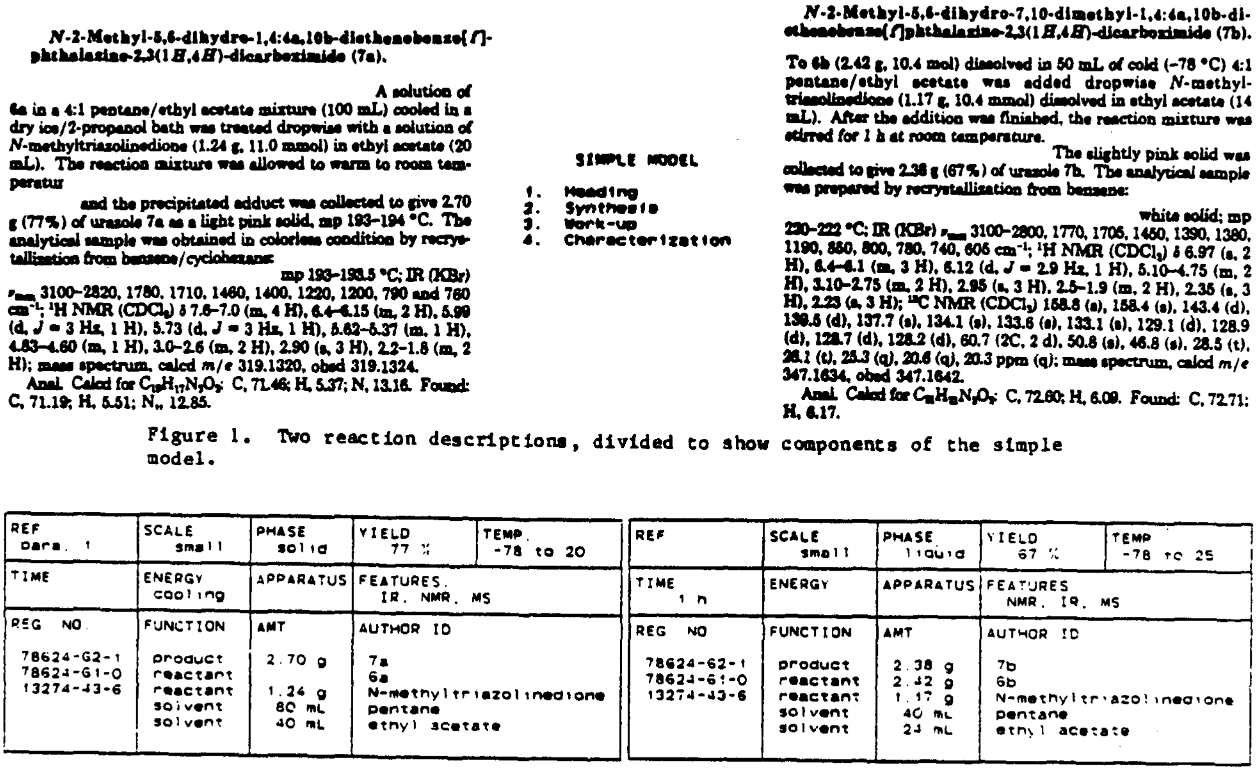
In the lexical stage, both dictionary lookup and morphological analysis are used to classify words. Morphological analysis procedures include suffix normalization, stemming and root word lookup and analysis of internal punctuation. Chemical substances may be identified by complex words and phrases, and are therefore surprisingly difficult to isolate.

Both lexical and syntactic means are used to isolate and tag chemical names. In the lexical stage, identifiable chemical roots, such as "benz" and terms, such as "iso-" are tagged. In the syntactic stage, a procedure uses clues such as parenthetical expressions, internal commas and the occurrence of juxtaposed chemical roots to identify chemical names. This is really morphology, of course. It also uses the overall syntax of the sentence to check whether a substance name is expected and to delimit the chemical name.

3. The Syntactic Stage

Chemical substances which comprise the reactants and the products of a chemical reaction, as well as the reaction conditions and yield, are identified by a hierarchical application of procedures. The syntactic stage of the system has been implemented by application of word expert procedures to the data structures built during the lexical stage.

The word experts are based upon the ideas of Rieger and Small (1979) but it has not been found to be necessary to use the full complexity of their model, so this system's word experts have



N-2-Methyl-5,6-dihydro-1,4:4a,10b-diethenobenzo[*f*]phthalazine-2,3(1*H*,4*H*)-dicarboximide (7a).

A solution of 6a in a 4:1 pentane/ethyl acetate mixture (100 mL) cooled in a dry ice/2-propanol bath was treated dropwise with a solution of *N*-methyltriazolinedione (1.24 g, 11.0 mmol) in ethyl acetate (20 mL). The reaction mixture was allowed to warm to room temperatur

and the precipitated adduct was collected to give 2.70 g (77%) of urazole 7a as a light pink solid, mp 193–194 °C. The analytical sample was obtained in colorless condition by recrystallization from benzene/cyclohexane:

mp 193–193.5 °C; IR (KBr) ν_{max} 3100–2820, 1780, 1710, 1460, 1400, 1220, 1200, 790 and 760 cm^{-1}; ^{1}H NMR ($CDCl_3$) δ 7.6–7.0 (m, 4 H), 6.4–6.15 (m, 2 H), 5.99 (d, J = 3 Hz, 1 H), 5.73 (d, J = 3 Hz, 1 H), 5.62–5.37 (m, 1 H), 4.83–4.60 (m, 1 H), 3.0–2.6 (m, 2 H), 2.90 (s, 3 H), 2.2–1.8 (m, 2 H); mass spectrum, calcd *m/e* 319.1320, obsd 319.1324.

Anal. Calcd for $C_{19}H_{17}N_3O_2$: C, 71.46; H, 5.37; N, 13.16. Found: C, 71.19; H, 5.51; N,, 12.85.

SIMPLE MODEL

1. Heading
2. Synthesis
3. Work-up
4. Characterization

N-2-Methyl-5,6-dihydro-7,10-dimethyl-1,4:4a,10b-diethenobenzo[*f*]phthalazine-2,3(1*H*,4*H*)-dicarboximide (7b).

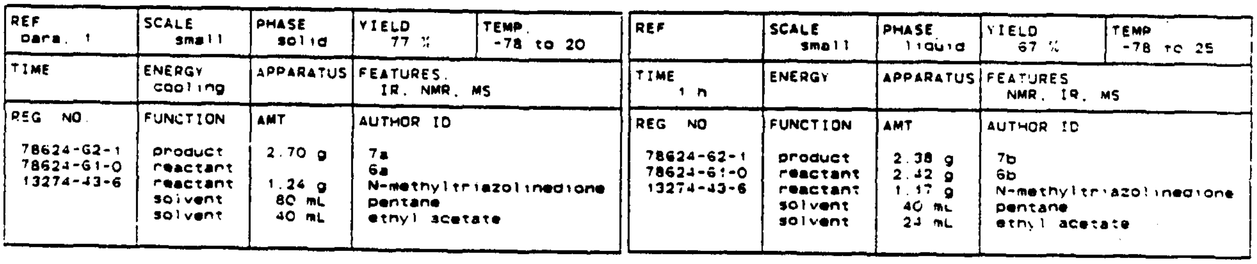
To 6b (2.42 g, 10.4 mol) dissolved in 50 mL of cold (−78 °C) 4:1 pentane/ethyl acetate was added dropwise *N*-methyltriazolinedione (1.17 g, 10.4 mmol) dissolved in ethyl acetate (14 mL). After the addition was finished, the reaction mixture was stirred for 1 h at room temperature.

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

white solid; mp 220–222 °C; IR (KBr) ν_{max} 3100–2800, 1770, 1705, 1450, 1390, 1380, 1190, 850, 800, 780, 740, 605 cm^{-1}; ^{1}H NMR ($CDCl_3$) δ 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); ^{13}C NMR ($CDCl_3$) 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), 28.1 (t), 25.3 (q), 20.6 (q), 20.3 ppm (q); mass spectrum, calcd *m/e* 347.1634, obsd 347.1642.

Anal. Calcd for $C_{21}H_{21}N_3O_2$: C, 72.60; H, 6.09. Found: C, 72.71; H, 6.17.

Figure 1. Two reaction descriptions, divided to show components of the simple model.



REF para. 1	SCALE small	PHASE solid	YIELD 77 %	TEMP. -78 to 20
TIME	ENERGY cooling	APPARATUS	FEATURES. IR, NMR, MS	
REG NO.	FUNCTION	AMT	AUTHOR ID	
78624-G2-1	product	2.70 g	7a	
78624-G1-0	reactant		6a	
13274-43-6	reactant	1.24 g	N-methyltriazolinedione	
	solvent	80 mL	pentane	
	solvent	40 mL	ethyl acetate	

REF	SCALE small	PHASE liquid	YIELD 67 %	TEMP -78 to 25
TIME 1 h	ENERGY	APPARATUS	FEATURES NMR, IR, MS	
REG NO	FUNCTION	AMT	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. Two reaction information forms, produced (manually) from the descriptions of Figure 1.

turned out to resemble a standard procedural implementation (Winograd, 1971) (based mostly on particular words or word categories, however). Their function is to determine the role of a word taking lexical and syntactic context into consideration. The word expert approach was initially chosen because it enables the implementation of fragments of a grammar and does not require the development of a comprehensive grammar. Since irrelevant portions can be identified by reliable heuristics and eliminated, this attribute is particularly useful in the SIE context. The procedures also allow the incorporation of heuristics for isolating certain items of interest.

In this context, it might be maintained that the interface between the syntax and the semantic mapping is even less clean than in certain other systems. This is intentional. Because of the specialized nature of the process, we have implemented the "semantic counterpart of syntax" concept, as advocated by Thompson (1966), where we judged that it would not impair the generality of the system within the area of reaction descriptions. We have tried not to make decisions that would make it difficult to extend the system to descriptions of reactions that do not obey the "simple model". The advantages of this approach were discused in Section I.

The system pays particular attention to verb arguments, which are generally marked by prepositions This "case" type analysis gives pretty good direct clues to the function of items within the meaning representation. Sentence structure is relatively regular, though extraposed phrases and a few types of clauses must be dealt with. Fortunately, the results, in terms of function of chemicals and reaction conditions, are the same whether the verb form is in an embedded clause or the main verb of the sentence. In other words, we do not have to deal with the nuances implied by higher predicates, or with implicative verbs, presuppositions, and the like.

4. The Semantic Stage

The semantic mapping could be directly to the components of the reaction information form, and that is the approach that was implemented in the first programs. This gave reasonable results in some test cases, but appeared to be less extensible to other models of reaction description than was desirable. A SNOBOL4 version maps the syntax to a predicate-argument formalism, with a case frame for each verb designating the possible arguments for each predicate.

5. The Extraction Stage

The meaning representation gives a pretty clear indication of the function of items within the RIF in the simple model. Since we wanted to experiment with generality in this system, we wished to separate general knowledge from linguistic knowledge, and for that reason, the actual extraction of items is done using the frame technique (Minsky, 1975; Charniak, 1975).

In the literature, frames and similar devices vary both in their format and in their function. In some cases, the information that they encode is still linguistic, at least in part. We are using them in the "nonlinguistic" sense, as discussed in Section I. In our system, frames encode the expectations that a trained reader would bring to the task of extracting information from synthetic descriptions, involving the usual structure of these descriptions.

A frame is being developed initially for the simple model. This frame looks for the synthesis section, discarding work-up and characterization except for the yield, which is usually to be found in the work-up. It then focuses on the synthesis, where subframes correspond to the particular entries needed in the RIF.

As one example, the "time" frame expects to find a series of reaction step times in the description. These are already labelled "time", and the frame will know that it has to total them, making approximations of such time expressions as "overnight" and indicating that the total is then approximate. Another example is the "temperature" frame, which expects a series of temperatures, and must calculate the minimum and maximum, in order to specify a range. Again, a certain amount of specialized knowledge, such as the temperature indicated by an ice water bath, is necessary.

C. Evaluation of the System

As of the date of this paper, we have only experimented with the version of the system that maps directly from the syntax into components of the reaction coding form. As noted above, this version does not have the generality that we desire, but gives a pretty good indication of the capabilities of the system, as now implemented.

As a test of the system, we ran it on fifty synthetic paragraphs from the experimental sections of the Journal of Organic Chemistry, and thirty-six were processed satisfactorily. Four had clear, detectable problems, so the robustness was 92%, but the accuracy was only 78%, since ten of the paragraphs did not follow the simple model, and were nevertheless processed. Since these were full of errors, we did not try to compute a figure for average error rate.

Although the objective of building this experimental system was only to deal with the simple model, the exercise has made clear to us the importance of the safety factor in making a system such as this useful. We intend to continue work with the present system only for a few weeks, meanwhile considering the problems and promises of extending it.

IV. RELATION TO SOME OTHER SIE SYSTEMS

The problem that we have had concerning the safety factor is one that is likely to be found in any SIE system, but it is soluble we feel. Even though we have not completed work on this experimental system as of the time of writing this paper (we have found more syntactic and semantic procedures to be implemented), we already have ideas as to how to build in a better safety factor. Generally, these can be characterized as using some of the information that can be gleaned by a combination of linguistic and chemical knowledge which we had ignored as redundant. While it is redundant in "successful" cases, it produces conflicts in other cases, indicating that something is wrong, and that the document should be processed by hand.

If the safety factor can be improved, SIE systems offer a promising area of application of computational linguistics techniques. Clearly, nothing less than computational linguistics techniques show any hope of providing a reasonable safety factor - or ever adequate robustness and accuracy.

The promise of the SIE area has been recognized by other researchers. Systems that fall within this paradigm include one constructed by the Operating Systems Division of Logicon (Silva, Montgomery and Dwiggins, 1979), which aims to "model the cognitive activities of the human analyst as he reads and understands message text, distilling its contents into information items of interest to him, and building a conceptual model of the information conveyed by the message," in the area of missile and satellite reports and aircraft activities. Another project, at Rutgers University, involves the analysis of case descriptions concerning glaucoma patients (Ciesielski, 1979), and the most extensive SIE project, also in the medical area, is that of the group headed by Naomi Sager (1981) at New York University, and described in her book.

V. REFERENCES

Charniak E. (1975). Organization and Inference in a Framelike System of Common Sense Knowledge. In R. C. Schank and B. L. Nash-Webber, eds., Theoretical Issues in Natural Language Processing, Mathematical Social Sciences Board.

Charniak, E. (1981). The Case-Slot Identity Theory, Cognitive Science, 5, 285-292.

Ciesielski, V. B. (1979). Natural Language Input to a Computer-Based Glaucoma Consultation System, In Proceedings, 17th Annual Meeting of the Association for Computational Linguistics, pp. 103-107.

Minsky, M. (1975). A Framework for Representing Knowledge. In P. Winston, ed., The Psychology of Computer Vision, McGraw-Hill, New York.

Rieger, C., and S. Small (1979). Word Expert Parsing. In Proceedings, Sixth International Joint Conference on Artificial Intelligence, Tokyo, 1979.

Sager, N. (1981). Natural Language Information Processing, Addison-Wesley, Reading, Massachusetts.

Silva, G., C. Montgomery and D. Dwiggins (1979). An Application of Automated Language Understanding Techniques to the Generation of Data Base Elements, In Proceedings, 17th Annual Meeting of the Association for Computational Linguistics, pp. 95-97. (See also The LOGICON report "A Language Understanding System for Automated Generation of AIA Data Base Elements", January 1981.)

Thompson, F. B. (1966). English for the Computer. In Proceedings, Fall Joint Computer Conference, Spartan Books, Washington, D. C.

Winograd, T. (1972). Understanding Natural Language, Academic Press, New York.

Zamora E. M., and L. H. Reeker (1982a). Computational Linguistics Research Project (Automatic Reaction Coding From English Descriptions), Lexical Phase (Tasks 1, 2.1, 2.2, 2.3, 2.4), Chemical Abstracts Service, March, 1982.

Zamora, E. M., and L. H. Reeker (1982b). Computational Linguistics Research Project (Automatic Reaction Coding From English Descriptions), Syntactic Phase (Tasks 2.5, 2.6), Chemical Abstracts Service, July, 1982.