AD-A157 892 DEVELOPING A NATURAL LANGUAGE INTERFACE TO COMPLEX DATA (U) SRI INTERNATIONAL MENLO PARK CA ARTIFICIAL INTELLIGENCE CENTER G G HENDRIX ET AL. AUG 77 1/1

UNCLASSIFIED SRI-TN-152 DAAG29-76-C-0012 F/G 9/2 NL

END

(1)

August 1977

AD-A157 892

DEVELOPING A NATURAL LANGUAGE INTERFACE TO COMPLEX DATA

by

Gary G. Hendrix
Earl D. Sacerdoti
Daniel Sagalowicz
and
Jonathan Slocum

Artificial Intelligence Center

Technical Note 152

The work reported herein, other than the development of the LIFER system, was supported by the Advanced Research Projects Agency of the Department of Defense under contract DAAG29-76-C-0012 with the U. S. Army Research Office. Development of LIFER was conducted under SRI International's Internal Research and Development Program.

DEVELOPING A NATURAL LANGUAGE INTERFACE TO COMPLEX DATA

by Gary G. Hendrix, Earl D. Sacerdoti, Daniel Sagalowicz, and Jonathan Slocum

Artificial Intelligence Center
SRI International
Menlo Park, California 94025

This paper describes aspects of an intelligent interface that provides natural language access to a large body of data distributed over a computer network. The overall system architecture is presented, showing how a user is buffered from the actual data base management systems (DBMSs) by three layers of insulating components. These layers operate in series to convert natural language queries into calls to DBMSs at remote sites. Attention is then focused on the first of the insulating components, the natural language system. A pragmatic approach to language access that has proved useful for building interfaces to data bases is described and illustrated by examples. Special language features that increase system usability, such as spelling correction, processing of incomplete inputs, and run-time system personalization, are also discussed.

Additional keywords: LADDER (Language Access to Distributed Data with Error Recovery); LIFER (Language Interface Facility with Spelling and Recursion); INLAND (Informal Natural Language Access to Navy Data).

~~I INTRODUCTION~~

In dealing with a very large data base (VLDB), which is perhaps distributed among multiple computers with different data base management systems (DBMSs) on remote sites, a central problem faced by would-be users is that of formulating queries in terms communicable to the system.

It is usually the case that business executives, government officials, and other decision makers have a good idea of the kind of information residing in their data bases. Yet to obtain the answer to a particular question, they generally need to employ the services of a technician who works with the data base on a regular basis and who is thoroughly familiar with its file structure, the DBMSs on which it resides, how it is distributed among various computer systems, the coded field names for the data items, the kinds of values that different fields are expected to contain, and other idiosyncrasies.

The technician must understand the decision maker's question, reformulate it in terms of the data that is actually stored, plan a sequence of requests for particular items from particular files on particular computers, open connections with remote sites, build programs to query the remote systems using the primitives of the remote systems' DBMSs, monitor the execution of those programs, recover from errors, and correlate the results. This is a demanding, time-consuming and exacting task requiring much attention to detail. Escalated levels of sophistication are needed as the VLDB increases in size and complexity, and as it is distributed over a wider range of host computers.

With the goal of making large, distributed data bases directly available to decision makers (while freeing technicians from increasingly tedious details), a group of researchers at SRI International has developed a prototype system that, for many classes of questions, automates the procedures usually performed by technicians. Subsequent sections of this paper present an overview of this system and then concentrate on the particular problem of translating user queries from English into the terms of the data base. The other aspects of the LADDER system are presented in greater detail elsewhere in the literature [12] [15] [16]. The system was developed as a management aid to Navy decision makers, so examples throughout the paper will be drawn from the domain of Navy command and control.

II SYSTEM ARCHITECTURE

Our running demonstration system, called LADDER* (for Language Access to Distributed Data with Error Recovery) [15], consists of three major components that provide levels of buffering of the user from the underlying DBMSs. The LADDER user can think he is retrieving information from a "general information base" rather than retrieving specific items of data from a set of highly formatted, traditional data bases that are scattered across a computer network. The user provides a question about the information base in English; LADDER applies all the necessary information concerning the vocabulary and syntax of the question, the names of specific fields, how they are formatted, how they are structured into files, and even where the files are physically located to provide an answer.

LADDER's first component, called INLAND (for Informal Natural Language Access to Navy Data), accepts questions in a restricted subset of natural language, and produces a query or sequence of queries to the VLDB as a whole. The queries to the VLDB, as produced by INLAND, refer to specific fields, but make no commitment about how the information in the data base is broken down into files.

For example, INLAND translates the question "What is the length of the Kennedy?" into the query

((NAM EQ JOHN#F.KENNEDY) (? LENGTH)),

where LENGTH is the name of the length field, NAM the name of the ship name field, and JOHN#F.KENNEDY the value of the NAM field for the record concerned with the Kennedy.

----
* A glossary of system names is provided below.

Queries from INLAND are directed to the second component of LADDER, called IDA (for Intelligent Data Access) [16]. In general, a query to IDA is a command list of constraints (such as (NAM EQ JOHN#F.KENNEDY) or (* MAX LENGTH)) and requests for values of fields (such as (? LENGTH)). INLAND operates by building a (possibly null) fragment of a query to IDA for each lower-level syntactic unit in the English input. These fragments are combined as higher-level syntactic units are recognized. At the sentence level, the combined fragments are sent as a command string to IDA.

Employing a model of the structure of the VLDB, IDA breaks down a query against the entire VLDB into a sequence of queries against individual files. Linkages among the records retrieved are preserved so that appropriate answers to the overall query may be composed and returned.

For example, suppose that the data base consists of a single file whose records contain the fields

(NAM CLASS LENGTH).

Then, to answer the data base query issued above, IDA can simply create one file retrieval program that says, in essence, "For the ship record with NAM equal JOHN#F.KENNEDY, return the value of the LENGTH field." Suppose, however, that the data base is structured in two files, as follows:

SHIP: (NAM CLASS ...)
CLASS: (CLASSNAME LENGTH ...).

In this case the single query about the Kennedy's length must be broken into two file queries. These would say, first, "Obtain the value of the CLASS field for the SHIP record with NAM equal JOHN#F.KENNEDY." Then, "Find the corresponding CLASS record, and return the value of the LENGTH field from that record."* Finally, IDA would compose an answer that is relevant to the user's original query (i.e. it will return NAM and LENGTH data, suppressing the CLASS-to-CLASSNAME link).

In addition to planning the correct sequence of file queries, IDA must actually compose those queries in the language of the remote DBMSs. Currently the system accesses, on a number of different machines, a DBMS called the Datacomputer [6] [4], whose input language is called Datalanguage. IDA creates the relevant Datalanguage query by inserting field and file names into pre-stored templates. However, since the data base is distributed over several machines, the Datalanguage that IDA produces does not refer to specific files in specific directories on specific machines. It refers instead to generic files, files containing a specific kind of record. For example, the queries discussed above might refer to the SHIP file rather than file SHIP.ACTIVE in directory NAVY on machine CCA-2.

It is the function of the third major component of LADDER to find the location of the generic files and manage the access to them. To carry out this function, the third component, called FAM (for File Access Manager) [12] relies on a locally stored model showing where files are located throughout the distributed data base. When it receives a query expressed in generic Datalanguage, it searches its model for the primary location of the file (or files) to which it refers. It then establishes connections over the ARPANET to the appropriate computers, logs in, opens the files, and transmits the Datalanguage query, as amended to refer to the specific files that are being accessed. If, at any time, the remote computer crashes, the file becomes inaccessible, or the network connection fails, FAM can recover, and, if a backup file is mentioned in FAM's model of file locations, it can establish a connection to a backup site and retransmit the query.

The existing system, written in INTERLISP [18], can process a fairly wide range of questions against a data base consisting of some 14 files containing about 100 fields. Processing a typical question takes less than a second of cpu time on a DEC KL-10 computer. An annotated transcript of a sample session with the system is provided in the Appendix.

We emphasize that the three major components of LADDER each address separate portions of the data access problem. Although they have been designed to work in combination, each component is a separate, self-contained module that independently addresses one aspect of data access. For example, the virtual view of the data that IDA supports for its caller would be of value even without a natural language front end. Likewise, the general technology developed for natural language translation may be separated from the data access problem and applied in other domains.

III THE NATURAL LANGUAGE COMPONENT

With the goal of supplying natural language interfaces to a variety of computer software, we have developed a language processing package called LIFER (for Language Interface Facility with Ellipsis and Recursion) [10] that facilitates the construction and run-time operation of special-purpose, applications-oriented, natural language interfaces. INLAND, the linguistic component of our intelligent interface to distributed data, has been constructed within the LIFER framework. Figure 1 gives some indication of the diversity of language accepted by this system. Below we describe the nature of INLAND and illustrate how it was created using LIFER's interactive language definition facilities. The examples, of course, can show only limited aspects of INLAND. We believe the existing INLAND system to be one of the most robust computerized natural language systems ever developed, accepting a wide range of questions about information in the data base (as shown in Figure 1) as well as metaquestions about definitions of data base fields and the grammar itself.

----
* If it is possible to perform multiple file accesses with a single multi-file query, IDA will do so.

FIGURE 1: A SAMPLE OF ACCEPTABLE INPUTS TO LADDER

What kind of information do you know about

Is there a doctor on board the Biddle

Display all the American cruisers in the North Atlantic

What is the name and location of the carrier nearest to New York

What is the commanding officer's name

Who commands the Kennedy

What is the Kennedy's beam

When will the Los Angeles reach Norfolk

Tell me when Taru is scheduled to leave port

Where is she scheduled to go

When will Los Angeles arrive in its home port

When will the Sturgeon arrive on station

What aircraft units are embarked on the Constellation

To which task organization is Knox assigned

Where is the Sellers

Where is Luanda

What is the next port of call of the Santa Inez

When will Tarifa get underway

Which convoy escorts have inoperative sonar systems

When will they be repaired

Which US Navy DDGs have casreps involving radar systems

What Soviet ship has hull number 855

To what class does the Soviet ship Minsk belong

What class does the Whale belong to?

What is the normal steaming time for the Wainwright from Gibraltar to Norfolk

What American ships are carrying vanadium ore

How far is it to Norfolk

How far away is Norfolk

How many nautical miles is it to Norfolk

How many miles is it to Norfolk from here

How close is the Baton Rouge to Norfolk

How far is the Adams from the Aspro

What is the distance from Gibraltar to Norfolk

What is the nearest oiler

What is the nearest oiler to the Constellation

How far is it from Naples to 23-00N, 45-00W

What is the distance from the Kittyhawk to Naples

How long would it take the Independence to reach 35-00N, 20-00W

How long is the Philadelphia

How long would it take the Aspro to join Kennedy

What is the nearest ship to Naples with a doctor c board

What is the nearest USN ship to the Enterprise wit an operational air search radar

What is known about that ship

How many merchant ships are within 400 miles of th Hepburn

What are their identities and last reported locations

What cargo does the Pecos have

Who is CTG 67.3

What are the length, width, and draft of the Kitty Hawk

To whom is the Harry E Yarnell attached

What type ships are in the Knox class

Where are the Charles F. Adams class ships

What are their current assignments

What subs in the South Atlantic are within 1000 miles of the Sunfish

What is the Kittyhawk doing

How many USN asw capable ships are in the Med

Where are they

What are their current assignments and fuel states

What ships are NOT at combat readiness rating C1

When will Reeves achieve readiness rating C1

Why is Hoel at readiness rating C2

When will the sonar be repaired on the Sterett

What ships are carrying cargo for the United States

Where are they going

What are they carrying

When will they arrive

Where is Gridley bound

Which cruisers have less than 50 per cent fuel on board

Where are all the merchant ships

When will the Kitty Hawk's radar be up?

What ships are in the Los Angeles class

What command does Adm. William have

Under whose opcon is the Dale

Show me where the Kennedy is!

What ship has hull number 148?

What is the next port of call for the South Carolina?

Are doctors embarked in the Kawishiwi

What kind of cargo does the Francis McGraw have?

What air group is embarked in the Constellation?

What do you know about the employment schedule of the Lang?

Which systems are down on the Kitty Hawk

What ships in the Med have doctors embarked?

How many ships carrying oil are within 340 miles of Mayport?

What sub contacts are within 300 miles of the Enterprise?

List the current position and heading of the US Navy ships in the Mediterranean every 4 hours

What is the status of the Enterprise's air search radar?

Where is convoy NL53 going

What convoy is the Transgermania in

How many embarked units are in Constellation

What ships are in British ports

What U S ships are within 500 miles of Wilmington?

What US ships faster than the Gridley are in Norfolk

What is the fastest ship in the Mediterranean Sea

How close is that ship to Naples?

What is its home port

Print the American cruisers' current positions and states of readiness!

How is the Los Angeles powered

What ship having a normal cruising speed greater than 30 knots is the largest

Display the last reported position of all ships that are in the North Atlantic

When did the Endeavour depart the port of New York

What nationality is the ship with international radio call sign UAlD

What ports are in the data base

What merchant ships are enroute to New York and within 500 miles of the Saratoga

To what country does the fastest sub belong?

A. Overview of LIFER

Although work in artificial intelligence and computational linguistics has not yet developed a general approach to the problems of understanding English and other natural languages, mechanisms do exist for dealing with major fragments of language pertinent to particular application areas. The idea behind LIFER is to adapt existing computational linguistic technology to practical applications while investigating and extending the human engineering aspects of the technology. The LIFER system supplies basic parsing procedures and an interactive methodology needed by a system developer to create convenient interfaces (such as INLAND) in reasonable amounts of time. Certain user-oriented features, such as spelling correction, processing of incomplete inputs, and the ability of the run-time user to extend the language accepted by the system through the use of paraphrase, are also included in the LIFER package.

LIFER is composed of two basic parts: a set o interactive language specification functions and a parser. The language specification functions are used to define an application language, a subset o a natural language (e.g., English) that is appropriate for interacting with existing software such as a DBMS. Using this language specification the LIFER parser interprets natural language inputs, translating them into appropriate interactions with the application software.

Figure 2 shows simplified example interaction with the LIFER parser using the INLAND language specification. A sequence of complete examples is presented in the Appendix. The user of the system types in a question or command in ordinary English followed by a carriage return. The LIFER parser then begins processing the input. When analysis is complete, the system types "PARSED!" and invokes data base functions (IDA) to respond.

An important feature of the parser is an ability to process elliptical (incomplete) inputs. Thus, if the system is asked, as in question 1,

WHAT IS THE LENGTH OF THE CONSTELLATION

then the input

OF THE NAUTILUS

will be interpreted as WHAT IS THE LENGTH OF THE CONSTELLATION.

If a user misspells a word, LIFER attempts to correct the error, using the INTERLISP spelling corrector [18]. If the parser cannot account for an input in terms of the application language definition, error messages, such as that produced after question 6, are printed that indicate how much of the input was understood and that suggest means of completing the input.

Provision is included in INLAND for interfacing with LIFER's own language specification functions, making it possible for users to give natural language commands for extending the language itself. In particular, naive users may extend the language accepted by the system by employing easy-to-understand notions such as synonyms and paraphrases. This is illustrated by interactions 7 and 10.

In using LIFER to define a language for INLAND, we have followed the approach taken by most real-time language processing systems in embedding considerable semantic information in the syntax of the language. Such a language specification is typically called a "semantic grammar." For example, words like NAUTILUS and DISPLACEMENT are not grouped together into a single <NOUN> category. Rather, NAUTILUS is treated as a <SHIP-NAME> and DISPLACEMENT as an <ATTRIBUTE>. Similarly, very specific sentence patterns such as

WHAT IS THE <ATTRIBUTE> OF <SHIP>

are typically used instead of more general patterns such as

<NOUN-PHRASE> <VERB-PHRASE>.

For each syntactic pattern, the language definer supplies an expression for computing the interpretation of instances of the pattern. INLAND's expressions for sentence-level patterns usually invoke the IDA component to retrieve information from the distributed data base.

FIGURE 2 SIMPLIFIED INTERACTIONS WITH LADDER

```
1-What is the length of the Constellation
PARSED!
(LENGTH 1072 feet)

2-of the Nautilus
TRYING ELLIPSIS:  WHAT IS THE LENGTH OF THE
                  NAUTILUS
(LENGTH 319 feet)

3-displacement
TRYING ELLIPSIS:  WHAT IS THE DISPLACEMENT OF THE
                  NAUTILUS
(STANDARD-DISPLACEMENT 4040 tons)

4-length of the fastest American Nuclear sub
TRYING ELLIPSIS:  WHAT IS THE LENGTH OF THE FASTEST
                  AMERICAN NUCLEAR SUB
(LENGTH 360 feet NAM LOS ANGELES  SPEED 30.0 knots)

5-Who commands the Constellation
      SPELLING-->CONSTELLATION
PARSED!
(COMMANDER  CAPT J.ELLISON)

6-Who commands JFK
TRYING ELLIPSIS:  ELLIPSIS HAS FAILED
THE PARSER DOES NOT EXPECT THE WORD "JFK" TO FOLLOW
"WHO COMMANDS"
OPTIONS FOR NEXT WORD OR META-SYMBOL ARE:
<SHIP-NAME>

7-Define JFK to be like Kennedy
PARSED!
 .  {JFK is now a synonym for KENNEDY,
 .   which is a ship name}
 .

8-Who commands JFK   {that is, retry interaction 6}
PARSED!
(COMMANDER CAPT P.MOFFETT)

9-info JFK country
TRYING ELLIPSIS:  ELLIPSIS HAS FAILED
 .  {error message omitted}
 .
10-Define "Info JFK country" to be like
     "what is the country of JFK"
PARSED!
 .
 .
11-Info JFK country
PARSED!
(NATION USA)

12-Info fastest American nuclear submarine speed
PARSED!
(SPEED 30.0 knots   NAM LOS ANGELES)

13-Nautilus
TRYING ELLIPSIS:  INFO NAUTILUS SPEED
(SPEED 22 knots)
```

This method of language specification is easy to understand and easy to use. But, when pursued systematically, it allows languages of rather broad coverage to be defined, as indicated in Figure 1.

To provide a more detailed view of how LIFER has been employed to produce an efficient and effective language processing system, let us examine in detail a highly simplified fragment of the INLAND language specification.

B. INLAND's Function in Brief

The central notions of how INLAND is constructed may be seen by considering the problem of providing English access to two files of the form

```
SHIP:(NAM CLASS COMMANDER HOME-PORT HULL# LOC)
CLASS: (CLASSNAME TYPE NATION FUEL LENGTH
                                  BEAM DRAFT SPEED)
```

located on different computers. IDA and FAM together provide levels of insulation from the real situation, so that INLAND need consider only the problem of specifying what subset of the overall data base should be queried, and what field values within that subset should be returned. IDA will dynamically plan the appropriate joins on the files in the data base, and FAM will carry them out. (In the actual LADDER system, the intertwining of multiple files is much more complex than in the current example.)

C. A Miniature Language Specification

1. Productions

The grammar rules may be viewed as productions of the form

metasymbol => pattern | expression,

where metasymbol is a metasymbol of the application language, pattern is a list of symbols and metasymbols in the language, and expression is a LISP expression whose value, when computed, is assigned as the value of the metasymbol. The symbol <L.T.G> (LIFER Top Grammar) is the highest-level metasymbol of the grammar. The system's answer to complete inputs that match a pattern instantiating <L.T.G> will be the result of evaluating the associated LISP expression.

For example, the input

PRINT THE LENGTH OF THE KENNEDY

is an instantiation of the sentence-level production

```
<L.T.G>  =>  <PRESENT> THE <ATTRIBUTE> OF <SHIP> |
                   (IDA (APPEND <SHIP> <ATTRIBUTE>)).
```

The input matches the pattern

<PRESENT> THE <ATTRIBUTE> OF <SHIP>,

where <PRESENT> matches PRINT, <ATTRIBUTE> matches LENGTH, and <SHIP> matches the phrase THE KENNEDY. If the semantic values for <SHIP> and <ATTRIBUTE>, computed by means to be described shortly, are ((NAM EQ JOHN#F.KENNEDY)) and ((? LENGTH)) respectively, then the answer to the question is computed from the expression portion of the production as follows:

```
    (IDA (APPEND <SHIP> <ATTRIBUTE>))
 => (IDA (APPEND '((NAM EQ JOHN.#F.KENNEDY))
                 '(? LENGTH)))
 => (IDA '((NAM EQ JOHN#F.KENNEDY)(? LENGTH))).
```

(APPEND is a LISP function that appends any number of lists together to form a larger list.) At this point, the IDA component is called with the argument

((NAM EQ JOHN#F.KENNEDY) (? LENGTH))

and the length of the Kennedy is retrieved:

```
    (IDA  '((NAM EQ JOHN F.KENNEDY) (? LENGTH)))
  => (LENGTH 1072 feet)
```

In LIFER, productions like the one shown above are defined interactively by issuing commands such as

```
PD[<L.T.G>
   (<PRESENT> THE <ATTRIBUTE> OF <SHIP>)
   (IDA (APPEND <SHIP> <ATTRIBUTE>))],
```

where PD is the production definition function.

2. Lexical Entries

Metasymbols, such as <PRESENT> and <ATTRIBUTE>, are often associated with individual words or fixed phrases, which are maintained in LIFER's lexicons. The LIFER function MS (Make Set) is used to define a set of words and phrases that may match a particular metasymbol. For example, the call

```
MS[<ATTRIB>
   (CLASS COMMANDER FUEL TYPE NATION LENGTH
   BEAM DRAFT (LOCATION . LOC) (POSITION . LOC)
   (NAME . NAM) (COUNTRY . NATION)
   (NATIONALITY . NATION)((HOME PORT) . HOME-PORT)
   ((POWER TYPE) . FUEL)((HULL NUMBER) . HULL#))]
```

is used to define sixteen words and fixed phrases that may match the symbol <ATTRIB> (which will be used subsequently in defining <ATTRIBUTE>).

After this call to MS, <ATTRIB> will match the words CLASS, COMMANDER, TYPE, FUEL, NATION, LENGTH, BEAM, and DRAFT. For these words, <ATTRIB> will take as its semantic value the word itself. <ATTRIB> will also match the word LOCATION, but for this match the value of <ATTRIB> will be the atom LOC. Similarly, <ATTRIB> matches POSITION, NAME, COUNTRY, and NATIONALITY but takes the values LOC, NAM, NATION, and NATION respectively. <ATTRIB> also matches the two-word phrase HOME PORT, taking HOME-PORT as its value. For the phrase POWER TYPE, the value is FUEL; for HULL NUMBER it is HULL#. (It is assumed that the codes HOME-PORT, HULL , LOC and NAM are peculiar to the data base and will not occur in natural language inputs.)

3. Subgrammars

Metasymbols may also be defined by production rules. For example, the call

```
PD[<ATTRIBUTE>
   (<ATTRIB>)
   (LIST (LIST '? <ATTRIB>))]
```

indicates that an <ATTRIBUTE> may be matched by an <ATTRIB>, viz:

```
          <ATTRIBUTE>  =>  <ATTRIB>.
```

For this production, the associated expression is

```
          (LIST (LIST '? <ATTRIB>)).
```

Since the word LENGTH matches <ATTRIB> and causes <ATTRIB> to take the atom LENGTH as its value, the rule above indicates that LENGTH is an instantiation of <ATTRIBUTE>. That is

```
   <ATTRIBUTE>  =>  <ATTRIB>  =>  LENGTH.
```

The value assigned to <ATTRIBUTE> when it matches LENGTH is computed by the production's expression as follows:

```
      (LIST (LIST '? <ATTRIB>))
   => (LIST (LIST '? 'LENGTH))
   => (LIST '(? LENGTH))
   => '((? LENGTH)).
```

This fragment of an IDA command requests the value of the LENGTH field. It was used above in answering the question "What is the length of the Kennedy?"

To recognize inputs such as
PRINT THE LENGTH BEAM AND DRAFT OF THE KENNEDY,*
the concept of an <ATTRIBUTE> may be generalized by adding two new productions as follows:

```
PD[<ATTRIBUTE>
   (<ATTRIB> AND <ATTRIBUTE>)
   (CONS (LIST '? <ATTRIB>) <ATTRIBUTE>)]

PD[<ATTRIBUTE>
   (<ATTRIB> <ATTRIBUTE>)
   (CONS (LIST '? <ATTRIB>) <ATTRIBUTE>)].
```

(CONS is a LISP function that adds an element (in this case the list whose first element is ? and whose second element is the value of <ATTRIB>) to the front of a list (in this case the value of <ATTRIBUTE>).) These productions allow the phrase LENGTH BEAM AND DRAFT to be accounted for in terms of the following syntax tree:

```
 LENGTH        BEAM       AND     DRAFT
    |            |         |        |
<ATTRIB>     <ATTRIB>      |     <ATTRIB>
    | V=LENGTH   | V=BEAM  |        | V=DRAFT
     \            \        |     <ATTRIBUTE>
      \            \       |      /  V=((? DRAFT))
       \             \     |     /
        \             <ATTRIBUTE>
         \           /  V=((? BEAM) (? DRAFT))
          \         /
          <ATTRIBUTE>
              V=((? LENGTH) (? BEAM) (?DRAFT))
```

4. Complete Analysis of a Simple Query

The examples above have indicated how the pattern

```
     <PRESENT> THE <ATTRIBUTE> OF <SHIP>
```

may be defined as a top-level input and how the metasymbol <ATTRIBUTE> may be defined. To complete the analysis of the top-level pattern, consider now the following definitions for <PRESENT> and <SHIP>.

To define <PRESENT>, the function MS may be used:

```
MS[<PRESENT>
   (PRINT LIST SHOW GIVE ((GIVE ME) . PRINT)
   ((WHAT IS) . PRINT) ((WHAT ARE) . PRINT))].
```

This call allows <PRESENT> to match the words PRINT, LIST, SHOW, and GIVE, and the phrases GIVE ME, WHAT IS, and WHAT ARE. The values assigned to <PRESENT>, which might be used, for example, to direct output to the terminal or to a graphics subsystem, are not of interest here.

----

* The use of two symbols <ATTRIB> and <ATTRIBUTE> could be avoided by letting <ATTRIBUTE> directly match lexical items and by introducing such productions as
<ATTRIBUTE> => <ATTRIBUTE> AND <ATTRIBUTE>
Unfortunately, the collapse of the two symbols into one results in both ambiguity and left recursion. LIFER recognizes only one of the ambiguous interpretations. Left recursion can be tolerated by special mechanisms in LIFER's top-down left-to-right parser, but only at a considerable increase in parsing time.

A <SHIP> may be designated in any one of a number of ways, the simplest being by name. The call

```
PD[<SHIP>
   (<SHIP-NAME>)
   (LIST (LIST 'NAM 'EQ <SHIP-NAME>))]
```

causes <SHIP> to match a <SHIP-NAME> and to take as its value an IDA command fragment restricting the value of the NAM field to be EQ (equal) to the particular name. <SHIP-NAME> may be defined by MS:

```
MS[<SHIP-NAME>
   (CONSTELLATION NAUTILUS
    (KENNEDY . JOHN#F.KENNEDY)
    ((JOHN F. KENNEDY) . JOHN#F.KENNEDY)  etc.)]
```

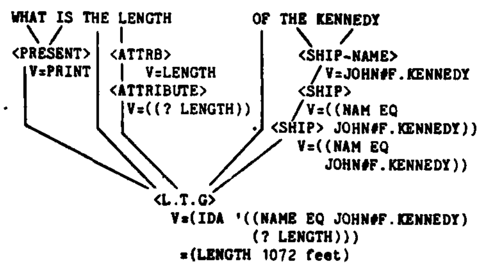
For an actual data base, this list is, of course, much more extensive. To allow the optional use of "the" before the name of a ship, a supplementary production for <SHIP> may be defined:

```
PD[<SHIP>
   (THE <SHIP>)
   <SHIP>].
```

With these definitions, LIFER has been given all the information needed to process a small class of sentence-level inputs. For example, the complete analysis of the input

WHAT IS THE LENGTH OF THE KENNEDY

is shown in the syntax tree



Note how the query given to IDA was generated by combining fragments from <SHIP> and <ATTRIBUTE>.

From the definitions for complete inputs defined above, LIFER can infer how to process incomplete inputs in context. For example, having just parsed the input

WHAT IS THE LENGTH OF THE KENNEDY

the system may, without additional knowledge, also handle the following sequence of incomplete inputs:

```
BEAM
 (i.e., what is the beam of the Kennedy)
HOME PORT AND CLASS
 (i.e., what is the home port
        and class of the Kennedy)
NAUTILUS
 (i.e., what is the home port
        and class of the Nautilus).
```

The method by which these incomplete inputs are processed is discussed below.

Other inputs that the rules defined thus far will accept include:

GIVE ME THE POSITION OF THE NAUTILUS
PRINT THE HULL NUMBER AND POWER TYPE
OF CONSTELLATION
SHOW THE COMMANDER COUNTRY AND TYPE
OF THE JOHN F. KENNEDY

D. Some Generalizations

The tiny fragment of language defined above already allows English access to most of the fields in the example data base, given the name of a ship. This fragment may be expanded easily along many dimensions.

1. Generalizing <SHIP>

Generalizing <SHIP> provides one of the most fruitful expansions. Naval ships are divided into major sets called classes. For example, the Constellation is in the Kitty Hawk class. Sometimes users will wish to ask questions about all ships of a particular class. To do this, the language may be extended by the call

```
PD[<SHIP>
   (<CLASS> CLASS SHIP)
   (LIST (LIST 'CLASS 'EQ <CLASS>))],
```

where <CLASS> is defined to match class names and takes their data base designations as values*. After this extension, the system will accept such inputs as

PRINT THE LENGTH OF KITTY HAWK CLASS SHIPS.

A <SHIP> might also match a general category such as CARRIERS, CRUISERS, or MERCHANT SHIPS. Such categories may usually be defined in terms of the TYPE field in the data base. For example, CARRIERS are of type CVA, CVAN, or CVS. OILERS are AO or AOR. <CATEGORY> might be defined by

```
MS[<CATEGORY>
   ((CARRIER . ((TYPE EQ CV)
                OR (TYPE EQ CVAN)
                OR (TYPE EQ CVS)))
    (OILER . ((TYPE EQ AO)
              OR (TYPE EQ AOR)))
    etc.)].
```

A new production for <SHIP> may then be added such as

```
PD [<SHIP>
    (<CATEGORY>)
    (LIST <CATEGORY>)].
```

With this production, the command

PRINT THE LOCATION OF CARRIERS

will be accepted.

Modifiers such as AMERICAN, NUCLEAR, and CONVENTIONAL are also very useful; for example,

```
MS[<MOD>
   ((AMERICAN . (NATION EQ US))
    (NUCLEAR . (FUEL EQ NUCLEAR))
    (CONVENTIONAL . (FUEL EQ DIESEL))
    etc.)].
```

By adding

```
PD[<SHIP>
   (<MOD> <SHIP>)
   (CONS <MOD> <SHIP>)],
```

the system will then process inputs such as

GIVE ME THE POSITION OF THE AMERICAN
NUCLEAR CARRIERS.

----

* To simplify the language definition, a language builder may supply LIFER with a preprocessor that does certain kinds of morphological transformations. For example, plural nouns such as SHIPS may be converted to the singular SHIP plus the pluralizing suffix -S. Or, as is assumed here, the suffix may simply be discarded.

Superlative modifiers, such as FASTEST and SHORTEST, may be defined:

```
MS[<MOD>
   ((FASTEST . (* MAX SPEED))
    (SLOWEST . (* MIN SPEED))
    (LONGEST . (* MAX LENGTH))
    etc.)].
```

Then the system will accept inputs such as

```
GIVE ME THE NAME AND LOCATION OF THE
  FASTEST AMERICAN OILERS.
```

This would translate into the IDA call

```
(IDA '((* MAX SPEED) (NATION EQ US)
       ((TYPE EQ AO) OR (TYPE EQ AOR))
       (? NAM) (? LOC))).
```

2. Generalizing <L.T.G>

New sentence-level productions, defined in terms of the more primitive metasymbols already described to LIFER, also greatly extend the range of language accepted. For example,

```
PD[<L.T.G>
   (<PRESENT> <SHIP>)
   (IDA (CONS '(? NAM) <SHIP>))]
```

allows inputs such as

```
  WHAT ARE THE FASTEST NUCLEAR SUBMARINES
  PRINT THE CARRIERS
and GIVE ME THE KITTY HAWK CLASS SHIPS.
```

As another example,

```
PD[<L.T.G>
   (WHO COMMANDS THE <SHIP>)
   (IDA (CONS '(? COMMANDER) <SHIP>))]
```

allows the input

```
WHO COMMANDS THE KENNEDY?
```

3. Calculated Answers

Sometimes a data base does not contain the information needed to answer a question directly, but nevertheless contains information that may be used as input to a procedure that can compute the answer from more primitive data. For example, the distance between two ships is not directly available in the example data base, although position data is. Suppose the function STEAMING.TIME can take speed and position information returned by IDA and calculate the time for the first ship to travel to the position of the second ship. Then, after defining <SHIP2> like <SHIP>,

```
PD[<SHIP2>
   (<SHIP>)
   <SHIP>],
```

a new top-level production may be defined as follows:

```
PD[<L.T.G>
   (HOW MANY HOURS IS <SHIP> FROM <SHIP2>)
   (STEAM.TIME
     (IDA (APPEND '((? SPEED)(? LOC)) <SHIP>))
     (IDA (CONS '(? LOC) <SHIP2>)))].
```

This production allows such queries as

```
HOW MANY HOURS IS KENNEDY FROM THE CONSTELLATION
```

E. Extending the Lexicon with Predicates

In certain instances, it is impractical to use the MS function to explicitly list all of the symbols that might match some metasymbol. For example, if the metasymbol <NUMBER> is to match any number, then MS is of little value. For such cases, LIFER allows a metasymbol to be associated with a predicate function. The metasymbol will match any symbol for which the predicate returns a non-NIL value. When such a match occurs, the metasymbol will take as its semantic value the response returned by the application of the predicate.

To define a metasymbol in terms of a predicate, the function MP (Make Predicate) is used. For example,

```
MP [<NUMBER>  NUMBERP]
```

defines <NUMBER> to match any symbol for which LISP predicate function NUMBERP returns a non-NIL value. When applied to numbers, NUMBERP returns the number itself. When applied to anything else, it returns NIL.

As the following questions indicate, <NUMBER> has many uses in the example data base:

```
WHAT CARRIERS HAVE LENGTHS GREATER THAN 1000 FEET
HOW FAR IS CONSTELLATION FROM 40 DEGREES NORTH 6
  DEGREES EAST
WHAT SHIPS ARE WITHIN 100 MILES OF KENNEDY.
```

As the size of the lexicon becomes large, the predicate feature may be used to push certain large classes of words out of the natural language system and into the data base itself. For example, <SHIP-NAME> could be defined in terms of a predicate that accesses the NAM field of the data base. (This would slow the parsing operation, of course, and spelling correction could not be performed easily.)

F. Accepting Metalanguage Inputs

It is possible to define input patterns that make reference to the LIFER package itself. For example, LIFER contains a function called SYMBOL.INFO which takes a metasymbol as its argument and prints lexical items, patterns and predicates that may be used to match the symbol. The interface builder may incorporate this function in response expressions as in

```
PD[<L.T.G>
   (HOW IS <SYMBOL> USED)
   (SYMBOL.INFO <SYMBOL>)]
```

After this call to PD, a user might ask the metaquestion

```
HOW IS <SHIP> USED
```

and receive the reply

```
<SHIP> MAY BE ANY SEQUENCE OF WORDS FOLLOWING
ONE OF THE PATTERNS:
  <SHIP> => <SHIP-NAME>
            THE <SHIP>
            <CLASS> CLASS SHIP
            <MOD> <SHIP>
```

LIFER also contains a function called PARAPHRASE that allows a new sequence of words to be defined as having the same meaning as a model sequence of words that the parser already accepts as a complete sentence. Using function PARAPHRASE in a response expression, the interface builder may extend the grammar by

```
PD[<L.T.G>
   (LET <SEQUENCE1> BE A PARAPHRASE OF <SEQUENCE2>)
   (PARAPHRASE <SEQUENCE1> <SEQUENCE2>)]
```

This new rule allows naive users to personalize the language understood by the system at run-time. For example, the user might say

```
  LET "REPORT ON KENNEDY" BE A PARAPHRASE OF
      "PRINT THE LOCATION AND COMMANDER OF KENNEDY"
```

The expression associated with the top-level production that matches this input sentence calls upon the paraphraser. Given the language definition defined above, LIFER then automatically adds the new production

```
          <L.T.G>  =>  REPORT ON <SHIP>
```

to the system, with an appropriate response expression. This new, user-defined production will allow the system to accept such new inputs as

```
     REPORT ON THE KENNEDY
     REPORT ON OILERS
     REPORT ON THE FASTEST AMERICAN SUBMARINES.
```

LIFER's methods for learning paraphrases are discussed below.

G. Extendibility

The subsections above have indicated how a few simple notions may be drawn together to create a small interface. But can the same notions be used to create much more sophisticated systems? Until our recent experience, we would have joined others in answering, "Not likely." Long before reaching an acceptable level of performance, previous language systems, including our own, have generally grown so complex and unwieldy that further extension has been stifled.

In designing LIFER, much attention has been given to the problem of supplying interface builders with an environment supporting the incremental development of relatively broad interfaces. All LIFER functions are interactive. Parsing and language specification tasks may be intermixed, allowing interface builders to operate in a rapid, extend-and-test mode. Transition trees, which are an efficient representation for the parser to work with, are automatically produced from productions, which we have found to be an efficient representation for interface builders to work with. The system contains a grammar editor and numerous special functions for answering questions about the structure of the language definition and for tracing and debugging a grammar. Details concerning these and other features of LIFER are specified more fully in the LIFER Manual [9].

We believe that the support features of LIFER have enabled us to give the INLAND language broader coverage than previous systems. Unfortunately, we know of no adequate measure of "breadth of coverage." However, some feeling for the types of inputs accepted by LADDER may be gained by considering a sample of acceptable inputs, such as that shown previously in Figure 1.

IV THE TRANSITION TREE PARSER

The LIFER parser is a top-down, left-to-right parser based on a simplification of Woods' [22] augmented transition network (ATN) system. Rather than use true ATNs, LIFER works with transition trees. IF <L.T.G> is defined by the productions

```
<L.T.G> => <PRESENT> THE <ATTRIBUTE> OF <SHIP> | e1
        => <PRESENT> <SHIP'S> <ATTRIBUTE> | e2
        => HOW MANY <SHIP> ARE THERE | e3
        => HOW MANY <SHIP> ARE THERE WITH
                                   <PROPERTY> | e4
```

then the following transition (not syntax) tree would be constructed for use by the parser:

```
                     THE-<ATTRIBUTE>-OF-<SHIP>--e1
                    /
          <PRESENT>-
         /          \
---------/           <SHIP'S>-<ATTRIBUTE>--e2
|<L.T.G>|
---------\                               e3
          \                             /
          HOW-MANY-<SHIP>-ARE-THERE
                                        \
             /--<-----<-----<----<--/
            \
             WITH-<PROPERTY>--e4
```

Starting at the box labled <L.T.G>, the parser attempts (nondeterministically) to move toward the response expressions on the right. At each step, the parser may move to the right on a branch if the left part of the remaining portion of the input can be matched by the symbol on the branch. Literal words on a branch can be matched only by themselves. A metasymbol, such as <PRESENT>, may be matched by a lexical item in the associated set created by MS. Or it may be matched by the predicate, if any, that has been defined for the metasymbol. Or it may be matched by successfully transversing some branch of the transition tree that encodes the productions expanding the metasymbol.

At the top level, if the parser reaches a response expression as a result of accounting for the last word of an input, then a top-level match for the input has been found and the response expression is evaluated to compute a response.

V IMPLEMENTATION OF SPECIAL LIFER FEATURES

This section presents an overview of LIFER's implementation of the spelling corrector, elliptical processor, and paraphraser.

A. IMPLEMENTATION OF SPELLING CORRECTION

LIFER uses a left-to-right parser based on a simplification of the ATN system of Woods [22]. Each time the parser discovers that it can no longer follow transitions along the current path, it records the failure on a failpoint list. Each entry on this list indicates the state of the system when the failure occurred (i.e., the position in the transition net and the values of various stacks and registers) and the current position in the input string. Local ambiguities and false paths make it quite normal for many failpoints to be noted even when a perfectly acceptable input is processed.

If a complete parse is found for an input, the failpoints are ignored. But if an input cannot be parsed, the list of failpoints is used by the spelling corrector, which selects those failpoints associated with the rightmost position in the input at which failpoints were recorded. It is assumed that failpoints occurring to the left were not caused by spelling errors, since some transitions using the words at those positions must have been successful for there to be failpoints to their right*.

The spelling corrector further restricts the rightmost failpoints by looking for cases in which a rightmost failpoint G is dominated by another rightmost failpoint F. G is dominated by F if G is a failpoint at the beginning of a subordinate transition tree that was reached in an attempt to expand F.

Working with the rightmost, dominating failpoints, the spelling corrector finds all categories of words that would be valid at the point where the suspected misspelling occurred. This typically requires an exploration of subgrammars. Using the INTERLISP spelling corrector, the word of the input string associated with the rightmost failpoints is compared with the words of the categories just found. If the "misspelled" word is sufficiently similar to any of these lexical items, the closest match is substituted. Failpoints associated with lexical categories that include the new word are then sequentially restarted until one leads to a successful parse. (This may produce more spelling correction further to the right.) If all restarts with the new word fail, other close lexical items are substituted for the "misspelled" word. If these also fail, LIFER prints an error message.

----

* This heuristic can cause LIFER to fail to find and correct certain errors. For example, if the user types CRAFT for DRAFT in WHAT DRAFT DOES THE ROARK HAVE, the spelling error will not be caught since a sentence such as WHAT CRAFT ARE NEAR ROARK would account for the initial sequence WHAT CRAFT. This is traded off against faster processing for the majority of spelling errors.

B. IMPLEMENTATION OF ELLIPSIS

LIFER's mechanism for treating elliptical inputs presumes that the application language is defined by a semantic grammar so that a considerable amount of semantic information is encoded in the syntactic categories. Thus, similar syntactic constructions are expected to be similar semantically. LIFER's treatment of ellipsis is based on this notion of similarity. During elliptical processing, LIFER is prepared to accept any string of words that is syntactically analogous to any contiguous substring of words in the last input. (If the last input was elliptical, its expansion into a complete sentence is used.)

LIFER's concept of analogy appeals to the syntax tree of the last input that was successfully analyzed by the system. For any contiguous substring of words in the last input, an "analogy pattern" may be defined by an abstraction process that works backwards through the old syntax tree from the words of the substring toward the root. Whenever the syntax tree shows a portion of the substring to be a complete expansion of a syntactic category, the category name is substituted for that portion. The analogy pattern is the final result after all such substitutions.

For example, consider how an analogy pattern may be found for the substring

OF SANTA INEZ,

using the syntax tree shown in Figure 3 for a previous input, WHAT IS THE LENGTH OF SANTA INEZ?

```
      WHAT IS THE   LENGTH    OF SANTA INEZ?
        \   /   |      |       |     \   /
     <PRESENT>  |   <ATTRIB>   |   <SHIP-NAME>
           \    |      |       |      /
            \   | <ATTRIBUTE>  | <SHIP>
             \   \        \    |  /
              \   \------\ \  /  /
               \          <ITEM>
                \         /
                  <L.T.G>
```

FIGURE 3: A Syntax Tree

Note that the syntax tree used here reflects production rules similar to those defined previously, but introduces a new metasymbol, <ITEM>, to add more substance to the discussion. Since the SANTA INEZ portion of the substring is a complete expansion of <SHIP-NAME>, the substring is rewritten as OF <SHIP-NAME>. Similarly, since <SHIP> expands to <SHIP-NAME>, the substring is rewritten as OF <SHIP>. Since no other portions of the substring are complete expansions of other syntactic categories in the tree, the process stops and OF <SHIP> is accepted as the most general analogy pattern. If the current input matches this analogy pattern, LIFER will accept it as a legitimate elliptical input. For example, the analogy pattern OF <SHIP>, extracted from the last input, may be used to match such current elliptical inputs as

OF THE KENNEDY
OF THE FASTEST NUCLEAR CARRIER
and OF KITTY HAWK CLASS SHIPS

Note that the expansion of <SHIP> need not parallel its expansion in the old input that originated the analogy pattern. For example, OF KITTY HAWK CLASS SHIPS is not matched by expanding <SHIP> to <SHIP-NAME> but by expanding <SHIP> to <CLASS> CLASS SHIP.

To compute responses for elliptical inputs matching OF <SHIP>, LIFER works its way back through the old syntax tree from the common parent of OF <SHIP> toward the root. First, the routine for computing the value of an <ITEM> from constituents of the production

<ITEM> => THE <ATTRIBUTE>

is invoked, using the new value of <SHIP> (which appeared in the current elliptical input) and the old value of <ATTRIBUTE> from the last sentence. Then, using the newly computed value for <ITEM> and the old value for <PRESENT>, a new value is similarly computed for <L.T.G>, the root of the syntax tree.

Some other substrings with their associated analogy patterns are shown below, along with possible new elliptical inputs matching the patterns.

substring:	THE LENGTH
pattern:	THE <ATTRIBUTE>
a match:	THE BEAM AND DRAFT
substring:	LENGTH OF SANTA INEZ
pattern:	<ATTRIBUTE> OF <SHIP>
a match:	HOME PORTS OF AMERICAN CARRIERS
substring:	WHAT IS THE LENGTH
pattern:	<PRESENT> THE <ATTRIBUTE>
a match:	PRINT THE NATIONALITY
substring:	WHAT IS THE LENGTH OF SANTA INEZ
pattern:	<L.T.G>
a match:	[any complete sentence]

For purposes of efficiency, LIFER's elliptical routines have been coded in such a way that the actual generation of analogy patterns is avoided*. Nevertheless, the effect is conceptually equivalent to attempting parses based on the analogy patterns of each of the contiguous substrings of the last input.

C. IMPLEMENTATION OF PARAPHRASE

LIFER's paraphrase mechanism also takes advantage of semantically oriented syntactic categories and makes use of syntax trees. In the typical case, the paraphraser is given a model sentence, which the system can already understand, and a paraphrase. The paraphraser's general strategy is to analyze the model sentence and then look for similar structures in the paraphrase string.

In particular, the paraphraser invokes the parser to produce a syntax tree of the model. Using this tree, the paraphraser determines all proper subphrases of the model, i.e., all substrings that are complete expansions of one of the syntactic categories listed in the tree. Any of these model subphrases that also appear in the paraphrase string are assumed to play the same role in the paraphrase as in the model itself. Thus, the semantically oriented syntactic categories that account for these subphrases in the model are reused to account for the corresponding subphrases of the paraphrase. Moreover, the relationship between the syntactic categories that is expressed in the syntax tree of the model forms a basis for establishing the relationship between the corresponding syntactic units inferred for the paraphrase.

a. Defining a Paraphrase Production

To find correspondences between the model and the paraphrase, the subphrases of the model are first sorted. Longer phrases have preference over shorter phrases, and for two phrases of the same length, the leftmost is taken first. For example, the sorted phrases for the tree of Figure 3 are

1.	<ITEM>	THE LENGTH OF SANTA INEZ	
2.	<PRESENT>	WHAT IS	
3.	<SHIP-NAME>	SANTA INEZ	--not used
4.	<SHIP>	SANTA INEZ	
5.	<ATTRIB>	LENGTH	--not used
6.	<ATTRIBUTE>	LENGTH	

Because the syntax tree indicates <SHIP> => <SHIP-NAME> => SANTA INEZ, both <SHIP-NAME> and <SHIP> account for the same subphrase. For such cases, only the most general syntactic category (<SHIP>) is considered. The category <ATTRIB> is similarly dropped.

Beginning with the first (longest) subphrase, the subphrases are matched against sequences of words in the paraphrase string. (If a subphrase matches two sequences of words, only the leftmost match is used.) The longer subphrases are given preference since matches for them will lead to generalizations incorporating matches for the shorter phrases contained within them. Whenever a match is found, the syntactic category associated with the subphrase is substituted for the matching word sequence in the paraphrase. This process continues until matches have been attempted for all subphrases.

For example, suppose the paraphrase proposed for the question of Figure 3 is

FOR SANTA INEZ GIVE ME THE LENGTH

Subphrases 1 and 2, listed above, do not match substrings in this paraphrase. Subphrase 3 is not considered, since it is dominated by subphrase 4. Subphrase 4 does match a sequence of words in the paraphrase string. Substituting the associated category name for the word sequence yields a new paraphrase string:

FOR <SHIP> GIVE ME THE LENGTH

Subphrase 5 is not considered, but subphrase 6 matches a sequence of words in the updated paraphrase string. The associated substitution yields

FOR <SHIP> GIVE ME THE <ATTRIBUTE>

Since there are no more subphrases to try, the structure

<L.T.G> => FOR <SHIP> GIVE ME THE <ATTRIBUTE>

----

* See Hendrix [10] for details of the algorithm.

is created as a new production to account for the paraphrase.

b. Defining a Response Expression for the Paraphrase Production

A new semantic response expression indicating how to respond to inputs matching this paraphrase production is programmed automatically from information in the syntax tree of the model. In particular, the syntax tree indicates which productions were used in the model to expand various syntactic categories. Associated with each of these productions is the corresponding response expression for computing the interpretation of the subphrase from subphrase constituents. The paraphraser reuses selected response expressions of the model to create a new expression for the paraphrase production. The evaluation of this new expression produces the same effect that would be produced if the expressions of the model were reevaluated. Metasymbols that appear in both the paraphrase production and the model remain as variables in the new response expression. Those symbols of the model that do not appear in the paraphrase production are replaced in the expression by the the constant values to which they were assigned in the model.

VI DISCUSSION

As implied by Figure 1 and the examples of the appendix, the INLAND system is a habitable, rather robust, real-time interface to a large data base and is fully capable of successfully accepting natural language inputs from inexperienced users. In the sections above, we have indicated some of the key techniques used in creating this system. We now seek to place our previous remarks in perspective by considering some of the limitations of the system, the roles played by the nature of our task and the tools we built in developing the system, and some of the similarities and differences between other systems and our own.

A. Limitations

In considering the limitations of our system, the reader should distinguish between limitations in the current INLAND grammar, which may easily be extended, and limitations in the underlying LIFER system, which is more or less fixed.

1. Syntactic Limitations

a. The Class of Languages Covered by LIFER

Consider the set of sentences that LIFER can accept. Because, in the worst case, a special top-level production may be defined in LIFER to cover any (finite-length) sentence that an interface builder may wish to include in the application language, it is impossible to exhibit a single sentence that the LIFER parser cannot be made to accept. Therefore, the only meaningful questions concerning syntactic limitations of LIFER must relate to LIFER's ability to use limited memory in covering infinite or large finite sets of sentences.

LIFER application languages are specified by augmented context free* grammars. Each rule in the grammar, as discussed previously, includes a context free production, plus an arbitrarily complex response expression, which is the "augmentation". Although a purely context free system would severely restrict the set of (non-finite) languages that LIFER could accept, the use of augmentation gives the LIFER parser the power of a Turing machine. The critical question is whether or not the context free productions and their more powerful augmentations can be made to support one another in meaningful ways.

To see the interplay between augmentation and context free rules in the recognition of a classic example of non-context free languages, consider the language composed of one or more X's followed by an equal number of Y's followed by an equal number of Z's. Let <x> be defined as

```
<x>  =>  X       | 1
<x>  =>  X <x>   | (PLUS 1 <x>)
```

Thus, <x> matches an arbitrary sequences of X's and takes as its value the number of X's in the string. Similar definitions may be made for <y> and <z>. A top-level sentence may be defined by the pattern <x><y><z>, but the augmentation must check to see that the numeric values assigned to the metasymbols are all equal. If they are equal, the augmentation expression returns some appropriate response. But if they are unequal, the expression returns the special atom *ERROR*, which the LIFER parser traps as a "semantic" (as opposed to syntactic) rejection.

The Turing machine power of LIFER is illustrated by the following trivial grammar.

```
<PRE-SENTENCE> => <WORD> | (LIST <WORD>)

               => <WORD> <PRE-SENTENCE>
                         | (CONS <WORD>
                                 <PRE-SENTENCE>)

<SENTENCE>     => <PRE-SENTENCE>
                         | (TMPARSE <PRE-SENTENCE>)
```

This grammar simply collects all of the words of the input into a list which is then passed to function TMPARSE, a parser of Turing machine power. In this extreme case, the LIFER parser makes virtually no use of the context free productions, but relies exclusively on the augmentation. LIFER is best used in the middle ground between this extreme and a purely context free system.

In other words, the class of languages for which LIFER was designed may be characterized as those allowing much of their structure to be defined by context free rules but requiring

----

* See Hopcroft and Ullman [11] for definitions of terms such as "context free" and "context sensitive".

occasional augmentation. It has been our experience that much of the subset of English used for asking questions about a command and control data base falls in this class. However, we have not considered certain complex types of transformations which will be discussed in the next subsection.

b. Troublesome Syntactic Phenomena

English speakers and writers often omit from a sentence a series of words that do not form a complete syntactic unit. For example, consider the following family of conjunctive sentences:

(1) WHAT LAFAYETTE AND WASHINGTON CLASS SUBS ARE WITHIN 500 MILES OF GIBRALTAR

(2) WHAT LAFAYETTE CLASS AND WASHINGTON CLASS SUBS ARE WITHIN 500 MILES OF GIBRALTAR

(3) WHAT LAFAYETTE CLASS SUBS AND KITTY HAWK CLASS CARRIERS IN THE ATLANTIC ARE WITHIN 500 MILES OF GIBRALTAR

(4) WHAT LAFAYETTE CLASS SUBS IN AND PORTS ON THE ATLANTIC ARE WITHIN 500 MILES OF GIBRALTAR

(5) WHAT LAFAYETTE CLASS SUBS IN THE ATLANTIC AND KITTY HAWK CLASS CARRIERS IN THE MEDITERRANEAN SOON WILL BE WITHIN 500 MILES OF GIBRALTAR

Sentence 1 omits the fragment CLASS SUBS ARE WITHIN 500 MILES OF GIBRALTAR from the "complete" question WHAT LAFAYETTE CLASS SUBS ARE WITHIN 500 MILES OF GIBRALTAR AND WHAT WASHINGTON CLASS SUBS ARE WITHIN 500 MILES OF GIBRALTAR. Note that the omitted fragment does not correspond to any well-formed syntactic unit, but begins in the middle of the noun phrase WHAT LAFAYETTE SUBS and continues to its right. Moreover, the fragment of the noun phrase that is left behind, namely WHAT LAFAYETTE, is not likely to be a well-formed syntactic unit, because one would expect to have WHAT combine with LAFAYETTE-SUBS rather than have WHAT-LAFAYETTE combine with SUBS. As the family of sentences above illustrates, the omission of words, signalled by the conjunction AND, may be moved to the right through the sentence one word at a time, slicing up the well-formed syntactic units at arbitrary positions. INLAND has no difficulty in accepting either conjunctions or disjunctions of well-formed syntactic categories, but LIFER provides no general mechanism for dealing with omissions that slice through categories at arbitrary points.

In the SYSCONJ system of Woods (1973), special mechanisms for handling a large (but not exhaustive) class of conjunction constructions were built into the parser. But conjunctions are just one example of the general need to create new "transformational" grammar rules at parse time.

A similar phenomenon occurs with comparative clauses, but much more is omitted and transformed. For example,
THE KITTY HAWK CARRIES MORE MEN THAN THE WASHINGTON may be viewed as a transformed and condensed form of

THE KITTY HAWK CARRIES X-MANY MEN AND
THE WASHINGTON CARRIES Y-MANY MEN AND
X IS MORE THAN Y.

For further discussion of this subject, see Paxton [14].

c. YES/NO Questions

A limitation of INLAND, although not of LIFER, is that few YES/NO questions are covered. The reason for this is pragmatic -- INLAND users do not ask them. Upon reflection, the motivation for this is clear -- WH questions produce more information for the questioner at a lower cost. A user might ask

IS THE KENNEDY 1000 FEET LONG

but it is shorter to ask

HOW LONG IS THE KENNEDY

and if the answer to the first question is NO (and if the system is so inconsiderate as to not indicate the correct length), then the user may have to ask for the length.

Creating a grammar for YES/NO questions is easy enough. For example,

```
PD[<L.T.G>
   (IS <NUMBER> <UNIT> THE <ATTRIB> OF <SHIP>)
   (YESNO.NUM.ATT <NUMBER>
     <UNIT> <ATTRIB> <SHIP>)]
```

might be used to allow the input

IS 1000 FEET THE LENGTH OF THE KENNEDY.

Function YESNO.NUM.ATT finds the <ATTRIB> of <SHIP> using IDA. Knowing the units in which he data base stores values of <ATTRIB>, YESNO.NUM.ATT converts the returned answer into the units specified by <UNIT> and compares the converted value to <NUMBER>. If the units are valid and the numbers match, YES is returned, otherwise NO is returned and the correct answer, as computed by IDA, is printed.

d. Assertions

INLAND was designed for retrieval and therefore does not handle such inputs as
THE LENGTH OF THE KENNEDY IS 1072 FEET.
LET THE LENGTH OF THE KENNEDY BE 1072 FEET.
SET THE LENGTH OF THE KENNEDY TO 1072 FEET.
Moreover, IDA does not provide for updating the data base.

However, let us assume that a command to IDA such as
((NAM EQ JOHN#F.KENNEDY)(! LENGTH 1072))
will cause the LENGTH field to be set to 1072 for the ship whose NAM field is EQ to JOHN#F.KENNEDY. Ignoring information about units and the problems of security and consistency, the following call to PD defines a class of update assertions.

```
PD[<L.T.G>
   (SET THE <ATTRIBUTE> OF <SHIP> TO <VALUE>)
   (IDA (APPEND <SHIP>
          (MAKE.ASSERTION
            <ATTRIBUTE> <VALUE>))))]
```

From two lists of equal length, such as
((LENGTH ?) (NATION ?))
and (1072 USA)

the function MAKE.ASSERTION creates a list of IDA assertion commands so that the IDA query commands on the first argument list will subsequently return the corresponding values on the second argument list. For the example shown, MAKE.ASSERTION would produce the assertion command list

```
((! LENGTH 1072) (! NATION USA))
```

The new production defined by PD will cover the inputs

```
SET THE LENGTH OF KENNEDY TO 1072
SET THE LENGTH AND COUNTRY OF KENNEDY
     TO 1072 AND USA
```

For the latter input, the actual call to IDA will be

```
(IDA '((NAM EQ JOHN F.KENNEDY)
       (! LENGTH 1072)
       (! NATION USA)))
```

A more complete response expression for this pattern would insure that the number of values equaled the number of attributes to be reset.

e. Irregular Coverage

One of the consequences of the ease with which interface builders can add new patterns to a LIFER grammar is that gaps may appear in coverage. For example, suppose a given language definition contains no passive constructions. Through the use of paraphrase or by direct action on the part of the interface builder, the language may be extended to cover some, but perhaps not all, passive constructions. That is, the system might be made to accept

1) THE KENNEDY IS OWNED BY WHOM

but not

2) THE KENNEDY IS COMMANDED BY WHOM.

(The semantically-oriented syntactic categories for OWNED and COMMANDED may differ.) If a user knows that the system accepts (1) and that the system accepts the active

3) WHO COMMANDS THE KENNEDY

then he is likely to be upset when input (2) is not accepted.

In creating the language specification for INLAND, we have tried by thoroughly modular programming to minimize such irregularities in coverage and have, we feel, been reasonably successful. Because LIFER gives the inference builder the freedom to add particular instances and subclasses of linguistic phenomena, it is his responsibility to avoid the gaps in coverage that may result.

2. Limitations Regarding Ambiguity

The LIFER parser does not deal with syntactic ambiguity directly, but accepts its first successful analysis as being the sole interpretation of an input. Because English contains truly ambiguous constructions, even when semantic considerations (the "augmentations") are taken into account, this limitation can be serious. For example, in the request

1) NAME THE SHIPS FROM AMERICAN HOME PORTS
   THAT ARE WITHIN 500 MILES OF NORFOLK

the phrase THAT ARE WITHIN 500 MILES OF NORFOLK might modify either the SHIPS or the PORTS. The choice will, of course, influence the response made to the user. The current LIFER parser is biased against deep parses and will only consider the interpretation in which the clause modifies SHIPS. Even a single word can produce difficulties. For example, the word NORFOLK in 1 could refer to a port in Virginia, a port in Great Britain, an American frigate, or a British destroyer. Thus, 1 is at least eight ways ambiguous.

Codd [3] has studied at some length the problem of ambiguity in the context of practical data base systems and has developed the strategy of engaging in a dialog in which the system articulates ambiguities (and other problems) and asks questions of the user to clarify the intent of his requests.

In addition to the simple syntactic form of ambiguity, exemplified by 1 above, other forms of ambiguity may arise. For example, the question

2) IS KENNEDY IN RADAR RANGE OF THE KNOX

is syntactically unambiguous, but the meaning might be either

IS KENNEDY IN KNOX'S RADAR RANGE
or IS KNOX IN KENNEDY'S RADAR RANGE.

This example represents a purely semantic ambiguity.

Section VI.A.1.b briefly touched on the need for transformational rules. As an example of how such rules might interact with the recognition of ambiguity, consider

3) IS KENNEDY NEARER TO GIBRALTAR THAN KITTY HAWK.

This question might be considered syntacticaly unambiguous in its present form, yet it has two possible meanings. By adding "missing words" at two different points in the input it is possible to produce the readings

IS THE KENNEDY NEARER TO GIBRALTAR THAN
  the kennedy is near to THE KITTY HAWK

and IS THE KENNEDY NEARER TO GIBRALTAR THAN
  THE KITTY HAWK is near to Gibraltar

Examples of ambiguity such as 1, 2, and 3 begin to show the difficulty of dealing with the problem in any general way. However, in the domain of INLAND, ambiguities have tended to arise only infrequently and have presented only minor problems for our particular application. Fortunately, our users have been very helpful by tending to avoid the use of the long and complex constructions that are most likely to lead to ambiguities.

Even though LIFER does not deal with ambiguity directly, certain types of ambiguities may be trapped and treated by using the response expressions of LIFER production rules. For example

```
PD[<L.T.G>
   (IS <SHIP1> IN <RANGE-TYPE> RANGE OF <SHIP2>)
   (COMPUTE.RANGE <SHIP1> <SHIP2> <RANGE-TYPE>)]
```

will accept such inputs as

IS KENNEDY WITHIN AIRCRAFT RANGE OF KNOX

and call the function COMPUTE.RANGE to respond. COMPUTE.RANGE is given the two ships and the range type as an input. Knowing the pattern to be inherently ambiguous, COMPUTE.RANGE may enter into a (formal) conversation with the user to resolve the ambiguity.

The INLAND grammar also tries to avoid ambiguity whenever possible. For example, the phrase AMERICAN ATTACK AIRCRAFT CARRIER might mean a ship that carries American attack aircraft, or an American attack ship that carries aircraft, or an American ship that carries attack aircraft, or any one of several other combinations. In INLAND, ATTACK-AIRCRAFT-CARRIER is recognized as a fixed phrase and all the problems with ambiguity vanish.

3. Limitations Regarding Definite Noun Phrases

INLAND has only a limited ability to handle definitely determined noun phrase such as THE SHIPS, THE AMERICAN SUB, and THOSE CRUISERS. As opposed to indefinite noun phrases (such as A SHIP), that refer to the existence of objects not currently in context, definite noun phrases are used to refer to a particular object or set of objects that is already in context. In dealing with a data base, "in context" may usually be taken to mean "in the data base." Thus, the phrase THE AMERICAN SUBS generally means "the American subs in the data base," and this is the interpretation that INLAND almost always places on this phrase. But suppose the user has just asked WHAT SUBS ARE IN THE MEDITERRANEAN and has been answered by a list of several subs, some of which are American and some of which belong to other countries. If the user now asks WHAT ARE THE POSITIONS OF THE AMERICAN SUBS he expects only the positions of American subs in the Mediterranean, but is given information about all American subs in the data base. The problem is that the local context established by previous questions is more restricted than the total data base and INLAND has not received enough clues to recognize this. (Had the input been WHAT ARE THE POSITIONS OF THE AMERICAN ONES the use of the pronoun would have signalled the local context and INLAND would have replied properly.)

Where the context is very clear, INLAND can sometimes handle a restricted perspective on the data base. For example, following

SELECT A MAP OF THE NORTH ATLANTIC

the query

DISPLAY THE AMERICAN SUBS

will cause the retrieval of only those subs in the North Atlantic, because others could not be displayed on the map in any case.

We know of no applied language system that deals adequately with this problem. However, significant experimental results are described in Grosz [ ].

4. Limitations in Processing Elliptical Inputs

After successfully processing the complete sentence

1) HOW MANY CRUISERS ARE THERE?

LIFER will accept the elliptical input

2) CRUISERS WITHIN 600 MILES OF THE KNOX

but not

3) WITHIN 600 MILES OF THE KNOX?

The elliptical processor is based on analogies. Input 2 is a noun phrase which is analogous to the noun phrase CRUISERS of input 1. Input 3, on the other hand, is a modifier that is intended to modify the CRUISERS of input 1. Because input 1 has no modifiers, elliptical input 3 has no parallel in the original input and hence cannot be accepted.

5. Other Limitations

A few other important limitations of INLAND and LIFER are worth mentioning briefly.

First, LIFER has no "core grammar" that is ready to be used on any arbitrary data base. This is because LIFER was designed as a general purpose language processing system and makes no commitment whatever to the types of programs and data structures for which it is to provide a front end or even to which natural language is to be accepted. This contrasts with systems such as Thompson's REL [19], which does provide a core grammar but which requires reformatting of data into the REL data base.

Some systems, such as ROBOT (Harris [8]), use the information in the actual data base as an extension of the language processor's lexicon. The LIFER interface may do this also but need not. If one elects not to use the data base as lexicon (and this choice was made in INLAND), then the lexicon must be extended whenever new values are added to the data base that a user may want to mention in his queries. The price of using the data base itself as an extended lexicon is that the data base must be queried during the parsing process. For very large data bases, this operation will probably be prohibitively expensive.

INLAND, of course, is basically a question answerer that relies on a data base as its major source of domain information. In particular, INLAND cannot read newspaper articles or other extended texts and record their meaning for subsequent querying. Moreover, although it is perfectly reasonable that the LIFER parser might be used for a text reading system, LIFER itself contains no particular facilities other than calls to response expressions for recording or reasoning about complex bodies of knowledge.

B. The Role of the Task Domain

The limitations presented in the last subsection would cause major difficulties in dealing with many areas of natural language application. However, for our particular application, the limitations did not prevent the creation of a robust and useful system. In the next few paragraphs, we briefly outline some of the key features that simplified our task.

The creation of INLAND was greatly facilitated by the nature of the particular interface problem that was addressed -- providing a decision maker with access to information he knows is in a data base. Because the user is expected to know what kinds of information are available and is expected to follow the technical terms and styles of writing that are typical in his domain of decision making, we can establish strong predictions about a user's linguistic behavior and hence INLAND needs to cover only a relatively narrow subset of language.

A second factor in facilitating the creation of the natural language interface was the interface provided by the IDA and FAM components of the LADDER system. By providing a simplistic view of what is in fact a complex and highly intertwined collection of distributed data, IDA and FAM helped greatly in simplifying the LISP response expressions associated with productions in the INLAND grammar.

In short, IDA allows the data base to be queried by high-level information requests that take the form of an unordered list of two kinds of items: fields whose values are desired, and conditions on the values of associated fields. Using IDA, the INLAND grammar need never be concerned with any entities in the data base other than fields and field values. Futhermore, because the input to IDA is unordered, the construction of segments of a call to IDA can be done while parsing lower-level metasymbols.

The performance of INLAND for a given user is also enhanced by the user's own, often unconscious, tendency to adapt to the system's limitations. Because INLAND can handle at least the most straightforward paraphrases of most requests for the values of any particular fields, even a new user has a good chance of having his questions successfully answered on the first or second attempt. It has been our experience that those who use the system with some regularity soon adapt the style of their questions to that accepted by the language specification. The performance of these users suggests that they train themselves to understand the grammar accepted by INLAND and to restrict their questions whenever possible to forms within the grammar. Formal investigation of this subjectively observed phenomenon might prove very interesting.

C. The Role of Human Engineering

Although the basic language processing abilities provided by LIFER are similar to those found in some other systems, LIFER embodies a number of human engineering features that greatly enhance its usability. These humanizing features include its novel ability to deal with incomplete inputs, and to allow naive users to extend the linguistic coverage at run time. But more importantly, LIFER provides easy-to-understand, highly interactive functions for specifying, extending, modifying, and debugging application languages. These features provide a highly supportive environment for the incremental development of sophisticated interfaces. Without these supporting features, a language definition rapidly becomes too complex to manage and is no longer extendable. With support, the relatively simple types of linguistic constructions accepted by LIFER may be used to produce far more sophisticated interfaces than was previously thought possible.

Creating a LIFER grammar that covers the language of a particular application may be thought of constructively as writing a program for a parser machine. All the precepts of good programming -- top-down design, modular programming and the like -- are relevant to good design of a semantic grammar. A well programmed grammar is easy to augment, because new top-level patterns are likely to refer to lower-level metasymbols that have already been developed and shown to work reliably. Thus, the task of adding new top-level productions to a grammar is analogous to the task of adding new capabilities to a more typical body of computer code (such as a statistics package) by defining new capabilities in terms of existing subroutines.

No matter how well-programmed a grammar might be, as the complexity of the grammar increases, the interactions among components of the language specification will grow. This leads the language designer into the familiar programming cycle of program, test, and debug. With many systems for parsing and language definition, the cycle may take many minutes for each iteration. With LIFER, when a new production is interactively entered into the grammar, it is immediately usable for testing by parsing sample inputs. The time required for the cycle of program, test, and debug is thus dependent on the thinking time of the designer, not the processing time of the system. Because the designer can make very effective use of his time, he can support, maintain, and extend a language specification of far greater complexity than would otherwise be possible.

The basic parsing technology of LIFER is not really new. But the human engineering that LIFER provides for interface builders has allowed us to better manage the existing technology and to apply it on a relatively large scale.

D. Related Work

As indicated by the February 1977 issue of the SIGART Newsletter [5], which contains a collection of 52 short overviews of various research efforts in the general area, interest in the development of natural language interfaces is widespread. Our own work is similar to that of several others.

The LIFER parser is a simplification of the ATN system developed by Woods [22]. Woods has used his parser in a variety of systems, including a speech understanding system, and a system [24] that uses a database in answering questions about the chemical analysis of lunar rocks. Wood's parser is also used by a number of other workers, notably Waltz [21] and Brown and Burton [1] [2]. The lunar rocks system does not use semantically oriented syntactic categories and the data base is smaller and less complex than that used by INLAND.

The first natural language systems to make extensive use of semantic grammars were those of Brown and Burton [1] [2]. These systems are designed for computer assisted instruction rather than as interfaces to data bases.

In work very similar to our own, Waltz [21] has devised a system called PLANES which answers questions about the maintenance and flight histories of airplanes. PLANES uses both an ATN and a semantic grammar. Apparently, the system does not include a paraphrase facility similar to LIFER'S. It does support the processing of elliptical inputs by a technique differing from our own, and supports clarification dialogues with users.

The PLANES language definition makes less use of syntactic information than INLAND. In particular, PLANES looks through an input for phrases matching certain semantically-oriented syntax categories. When one of these phrases is found, it is placed in a local register that is associated with the given category. Rather than attempt to combine these phrases into a complete sentence by syntactic means, "concept case frames" are used. Essentially, PLANES uses case frames to decide what type of question has been asked by looking at the types and values of local registers that were set by the input. For example, the three questions

WHO OWNS THE KENNEDY
BY WHOM IS KENNEDY OWNED
THE KENNEDY IS OWNED BY WHOM

would all set, say, an <ACT> register to OWN and a <SHIP> register to KENNEDY. The case frames can determine what question is asked simply by looking at these registers. Performing a syntactic analysis such as LIFER does requires different constructions for each question pattern.

If the next input is the elliptical fragment "KNOX", the <SHIP> register is reset. Because no case frame is associated with <SHIP> alone and because <SHIP> was used in the last input, the <ACT> register is inherited in the new context and the elliptical input properly analyzed. When more than one case frame matches an input, PLANES enters into a clarification dialog with the user to decide which was intended. (This conversation prints the data base calls and asks the user to choose among them.)

The use of case frames is very attractive in that it allows many top-level syntactic patterns to be accounted for by a single rule. However, it is inadequate for complex inputs. The question IS KNOX FASTER THAN KENNEDY contains two <SHIP>s. Only the syntax tells us which to test as the faster of the two. Compound-complex sentences would be extremely difficult to process without extensive use of syntactic data.

Codd's RENDEZVOUS system [3] for interface to relational data bases provides many ideas concerning clarification dialogs that might be included in LIFER at some later date. RENDEZVOUS is fail-safe in that it can fall back on multiple choice selection if natural language processing fails completely.

Another applied natural language system whose underlying philosophy is akin to that of LIFER is the REL (Rapidly Extendable Language) system of Thompson and Thompson [19]. REL is a data retrieval system like LADDER, though REL requires data to be stored in a special REL data base. The grammar rules of REL contain a context free part and an argumentation very much like those of LIFER. As its name implies, REL was intended to be easily extendable by interface builders. Much effort has gone into making REL run rapidly and it is almost certainly faster than LIFER. However, this speed was gained by a low level language implementation with the unfortunate side result that response expressions are not easily written.

Recently, the Artificial Intelligence Corporation introduced a commercial product called ROBOT for interfacing to data bases. As described in Harris [8], ROBOT "calls for mapping English language questions into a language of data base semantics that is independent of the contents of the data base." The data base itself is used as an extension of the dictionary and the structure of files within the data base helps in guiding the parser in the resolution of ambiguities. Our own research indicates that the types of linguistic construction used are rather dependent on the content of the data base. We also worry that extensive recourse to a data base of substantial size may greatly slow the parsing process. Moreover, our data base is coded largely in terms of abbreviations that are unsuitable as lexical entries. Nevertheless, the notion of using the data itself to extend the capabilities of the language system is very attractive.

In addition to the work on near-term application systems, a number of workers are currently addressing longer-range problems of accessing data bases through natural language. These include Mylopolous et al. [13], Sowa [17] Walker et al. [20], and ourselves [15]. There are, of course, many people engaged in research in the general area of natural language processing, but a survey of their work is beyond the scope of this paper.

VII CONCLUSION

We have described a system called LADDER that provides natural language access to a large, distributed data base. We have shown that the language processing component of this system, although based on simple principles and subject to certain limitations, is sufficiently robust to be useful in practical applications. Moreover, we have indicated that LADDER is not an isolated system but that other applied language systems have achieved significant levels of performance as well, particularly in interfacing to data bases. We believe that the evidence presented indicates clearly that, for certain restricted applications, natural language access to data bases has become a practical and practicable reality.

VIII ACKNOWLEDGEMENTS

The work reported herein, other than the development of the LIFER system, was supported by the Advanced Research Projects Agency of the Department of Defense under contract DAAG29-76-C-0012 with the U. S. Army Research Office. Development of LIFER was conducted under SRI

International's Internal Research and Development Program.

IX GLOSSARY

DBMS -- Data Base Management System

FAM -- File Access Manager. Maps generic file names onto specific file names on specific computers at specific sites. Initiates network connections, opens files, and monitors for certain errors.

IDA -- Intelligent Data Access. Presents a structure-free view of a distributed data base.

INLAND -- Informal Natural Language Access to Naval Data. The natural language interface to IDA, which incorporates a special-purpose LIFER grammar.

LADDER -- Language Access to Distributed Data with Error Recovery. Our total system composed of INLAND, IDA, and FAM.

LIFER -- Language Interface Facility with Ellipsis and Recursion. The general facility for creating and maintaining linguistic interfaces.

MS -- Make Set. The LIFER function for defining a metasymbol as a set of lexical items.

MP -- Make Predicate. The LIFER function for defining a metasymbol as a predicate function.

PD -- Pattern Define. The LIFER function for defining a metasymbol as a pattern expansion.

VLDB -- Very Large Data Base

REFERENCES

1. J. S. Brown and R. R. Burton, "Multiple Representations of Knowledge for Tutorial Reasoning," pp. 311-349, D. G. Bobrow and A. Collins, eds., Representation and Understanding (Academic Press, New York, 1975).

2. R. R. Burton, "Semantic Grammar: An Engineering Technique for Constructing Natural Language Understanding Systems," BBN Report No. 3453, Boston, Mass. (December 1976).

3. E. F. Codd, "Seven Steps to Rendezvous with the Casual User," pp. 179-200, J. W. Klimbie and K. I. Koffeman, eds., Data Base Management, (North-Holland, 1974).

4. Computer Corporation of America, "Datacomputer Version 1 User Manual," CCA, Cambridge, Mass. (August 1975).

5. L. D. Erman, ed., SIGART Newsletter, No. 61 (ACM, New York, February 1977).

6. J. Farrell, "The Datacomputer - A Network Data Utility," Proc. Berkeley Workshop on Distributed Data Management and Computer Networks, Berkeley, Ca., pp. 352-364 (May 1976).

7. B. J. Grosz, "The Representation and Use of Focus in Dialog Understanding," Ph.D. dissertation, U. C. Berkeley (May 1977).

8. L. R. Harris, "ROBOT: A High Performance Natural Language Processor for Data Base Query," pp. 39-40 in [5].

9. G. G. Hendrix, "The LIFER Manual: A Guide to BuildingPractical Natural Language Interfaces," SRI Artificial IntelligenceCenter Tech. Note 138, Menlo Park, Ca. (February 1977).

10. G. G. Hendrix, "Human Engineering for Applied Natural Language Processing," Proc. 5th International Joint Conference on Artificial Intelligence, Cambridge, Mass. (August 1977).

11. J. E. Hopcroft and J. D. Ullman, Formal Languages and their Relation to Automata (Addison-Wesley, Reading, Mass. 1969).

12. P. Morris and D. Sagalowicz, "Managing Network Access to a Distributed Data Base," Proc. Second Berkeley Workshop on Distributed Data Management and Computer Networks, Berkeley, Ca. (May 1977).

13. J. Mylopolous, A. Borgida, P. Cohen, N. Roussopoulos, J. Tsotsos, and H. Wong, "TORUS - A Natural Language Understanding System for Data Management," Proc. 4th International Joint Conference on Artificial Intelligence, Tbilisi, U.S.S.R. (August 1975).

14. W. H. Paxton, "A Framework for Speech Understanding," SRI Artificial Intelligence Center Tech. Note 142, Menlo Park, Ca. (June 1977).

15. E. D. Sacerdoti, "Language Access to Distributed Data with Error Recovery," Proc. 5th International Joint Conference on Artificial Intelligence, Cambridge, Mass, (August 1977).

16. D. Sagalowicz, "IDA: An Intelligent Data Access Program," Proc. Third International Conference on Very Large Data Bases, Tokyo, Japan (October 1977).

17. J. F. Sowa, "Conceptual Graphs for a Data Base Interface," IBM Journal of Research and Development, Vol. 20 No. 4, pp. 336-357 (July 1976).

18. W. Teitelman, "INTERLISP Reference Manual," Xerox PARC, Palo Alto, Ca. (December 1975).

19. F. B. Thompson and B. H. Thompson, "Practical Natural Language Processing: The REL System as Prototype," pp. 109-168, M. Rubinoff and M. C. Yovits, eds., Advances in Computers 13 (Academic Press, New York, 1975).

20. D. E. Walker, B. J. Grosz, G. G. Hendrix, W. H. Paxton, A. E. Robinson, and J. Slocum, "An Overview of Speech Understanding Research at SRI," Proc. 5th International Joint Conference on Artificial Intelligence, Cambridge, Mass. (August 1977).

21. D. Waltz, "Natural Language Access to a Large Data Base: an Engineering Approach," Proc. 4th Internatioal Joint Conference on Artificial Intelligence, Tbilisi, USSR, pp. 868-872 (September 1975).

22. W. A. Woods, "Transition Network Grammars for Natural Language Analysis," CACM, Vol. 13, No. 10, pp. 591-606 (October 1970).

23. W. A. Woods, "An Experimental Parsing System for Transition Network Grammars," in R. Rustin, ed., Natural Language Processing (Algorithmics Press, New York, 1973).

24. W. A. Woods, R. M. Kaplan, and B. Nash-Webber, "The Lunar Sciences Natural Language Information System," BBN Report 2378, Bolt Beranek and Newman, Cambridge, Mass. (1972).

APPENDIX

A TRANSCRIPT OF INTERACTIONS WITH LADDER

```
@ladder
Please type in your name: D. T. Base

Do you want instructions? (type FIRST LETTER of response) Yes

        This program has access to 14 files which constitute a facsimile
of a Navy command and control data base for the Atlantic and Mediterranean
areas. The data is stored at NOSC in San Diego, and on the Datacomputer at
CCA in Cambridge, Mass.  The data base includes physical characteristics and
position information for all ships, with more detailed operational information
about U. S. Navy ships.  Data about embarked USN units, convoys of merchant
ships, and ports of arrival and departure are also available.

        The system accepts queries stated in English, and will translate them
into calls to a DBMS. It will respond to the question  DESCRIBE THE FIELDS
with a short description of each of the 72 fields in the data base. For
further information about the contents of the data base, a Technical Note
is available from NOSC.

        IMPORTANT NOTE: the current version of the query answering system can
perform only simple calculations on the values in the data base. For example,
it cannot answer questions about the composition of organizational units.

Example questions include:
WHAT IS THE NEAREST SHIP TO LUANDA WITH A DOCTOR ABOARD?
HOW FAR IS THE CONSTELLATION FROM CHARLESTON?
WHERE IS THE LOS ANGELES?
THE LONGEST SHIP CARRYING VANADIUM ORE?
WHEN WILL THE PHILADELPHIA REACH PORT?
WHAT U. S. SHIPS ARE WITHIN 500 MILES OF GIBRALTAR?
WHAT PORTS DO YOU KNOW ABOUT?
WHAT IS THE STEAMING TIME FROM NORFOLK TO GIBRALTAR FOR THE STERETT?
CAN THE JFK REACH NORFOLK WITHOUT REFUELING?

Do you want to use 2 Data Computers? No
Do you want to specify a current location (default = Norfolk)? No

Do you wish distance/direction calculations to default to GREAT CIRCLE,
or RHUMB LINE? (you can override by specifying in the query) Great Circle

1_What is the current position of the Kennedy?
PARSED!
Parse time: .74 seconds
        * This counts cpu time used by INLAND.
 Connecting to Datacomputer at CCA1:
        * FAM indicates which computer is being accessed.  The next 12
        * lines are interactions between FAM and the Datacomputer.
>> ;O031 770620185344   IONETI: CONNECTED TO SRI-KL-22700010
>> ;J150 770620185413   FCRUN: V='DC-3/60.00.2' J=3 DT='MONDAY, JUNE 20, 1977 1
:54:13-EDT' S='CCA'
>> !0041 770620185414   DNCTNX: DATACOMPUTER GOING DOWN IN 2273 MIN BECAUSE SYS
EM IS GOING DOWN AT WED JUN 22 77 5:00:00AM-EDT FOR 240 MIN DUE TO SCHEDULED PM
>> ;J200 770620185414   RHRUN: READY FOR REQUEST
*> Set parameters
*<      Exit
CCA1:^Z
*> Set parameters
*< V    Verbosity (-1 to 5):   1
*< P    PROCEED with Datalanguage  [confirm with <CR>]
        * The connection has been established.  FAM now logs in and
        * opens the necessary files.
CCA1:LOGIN %TOP.ACCAT.GUEST ;
CCA1:OPEN %TOP.ACCAT.SAGALOWICZ.NSTDPORT1 WRITE;
CCA1:OPEN %TOP.ACCAT.NTRACKHIST READ;
CCA1:OPEN %TOP.ACCAT.NNSHIP READ;
        * FAM now transmits the query.
CCA1:FOR R1 IN NNSHIP WITH (NAM EQ 'KENNEDY JF') FOR NSTDPORT1 , R2 IN
CCA1:NTRACKHIST WITH R2.UICVCN EQ R1.UICVCN BEGIN STRING1 = R2.PTP STRING2 =
CCA1:R2.PTD END;
*> Total bytes transferred: 27
```

```
Computation time for query: 3.933 seconds
        * This counts cpu time used by IDA and FAM.
Real time for query: 571.485 seconds
        * This counts real time from the time the request to IDA is
        * made until IDA returns its answer.
(POSITION 6000N03000W DATE 7601171200)

2_of kitty hawk
 Trying Ellipsis:  WHAT IS THE CURRENT POSITION OF KITTY HAWK
Parse time: 1.494 seconds
CCA1:FOR R1 IN NNSHIP WITH (NAM EQ 'KITTY HAWK') FOR NSTDPORT1 , R2 IN
CCA1:NTRACKHIST WITH R2.UICVCN EQ R1.UICVCN BEGIN STRING1 = R2.PTP STRING2 =
CCA1:R2.PTD END;
*> Total bytes transferred: 27
Computation time for query: .604 seconds
Real time for query: 65.244 seconds
(POSITION 3700N01700E DATE 7601171200)

3_To what country does the fastest sub belong
PARSED!
Parse time: .501 seconds
CCA1:OPEN %TOP.ACCAT.SAGALOWICZ.NSTDPORT WRITE;
CCA1:OPEN %TOP.ACCAT.NNSHIPCLASDIR READ;
CCA1:OPEN %TOP.ACCAT.NNSHIPCLASCHAR READ;
CCA1:BEGIN DECLARE Z STRING (,100) ,D='}' DECLARE X STRING (,100) ,D='}' DECLAR
E
CCA1:Y STRING (,100) ,D='}' Y = '*' DECLARE Y1 STRING (,100) ,D='}' Y1 = '*'
CCA1:DECLARE Y2 STRING (,100) ,D='}' Y2 = '*' X = '00.0' FOR R1 IN
CCA1:NNSHIPCLASCHAR WITH (TYPE1 EQ 'S') AND (TYPE2 EQ 'S') FOR R2 IN
CCA1:NNSHIPCLASDIR WITH R2.SHIPCLAS EQ R1.SHIPCLAS FOR R3 IN NNSHIP WITH
CCA1:R3.UICVCN EQ R2.UICVCN BEGIN Z = R3.MCSF IF Z LT '99.9' AND X LT Z THEN
CCA1:BEGIN Y = R3.NAT Y1 = R3.NAM X = Z END END NSTDPORT . STRING1 = Y NSTDPORT
CCA1:. STRING2 = Y1 NSTDPORT . STRING3 = X END;

*> Total bytes transferred: 47
Computation time for query: 2.585 seconds
Real time for query: 440.771 seconds
(NAT US SHIP LOS ANGELES MXSPD 30.0)

4_each merchant ship in the North Atlantic
 Trying Ellipsis:  TO WHAT COUNTRY DOES EACH MERCHANT SHIP IN THE
        NORTH ATLANTIC BELONG
Parse time: 1.111 seconds
CCA1:OPEN %TOP.ACCAT.SAGALOWICZ.NSTDPORT2 WRITE;
CCA1:CLOSE NSTDPORT1 ;
CCA1:OPEN %TOP.ACCAT.NNMOVES READ;
CCA1:FOR R1 IN NNMOVES WITH ((TYPE EQ 'BULK') OR (TYPE EQ 'TNKR')) FOR R2 IN
CCA1:NTRACKHIST WITH (PTPNS EQ 'N') AND (PTPEW EQ 'W') AND
CCA1:((PTPY GT 600) OR (PTPX LT 3600) OR (PTPX GT 3900)) AND R2.UICVCN EQ
CCA1:R1.UICVCN FOR NSTDPORT2 , R3 IN NNSHIP WITH R3.UICVCN EQ R2.UICVCN BEGIN
CCA1:STRING1 = R2.PTP STRING2 = R2.PTD STRING3 = R3.NAT STRING4 = R3.NAM END;
*> Total bytes transferred: 4582
Computation time for query: 4.833 seconds
Real time for query: 441.251 seconds
        * When a large number of answers is returned, LADDER makes
        * sure the user really wants them.
I have 79 answers; do you want to see them all? Yes
NAT SHIP                 POSITION    DATE
US  GREENVILLE VICTORY 1533N03010W 7601171200
US  TOULE J              1533N03010W 7601171200
        .
        . {other entries omitted here}
        .
WG  TRANSQUEBEC          4633N02129W 7601171200
NO  TALABOT              0600N01130W 7601171200

5_what ships are en route to america
 Trying Ellipsis:  Ellipsis has failed.
THE PARSER DOES NOT EXPECT THE WORD "AMERICA" TO FOLLOW
"WHAT SHIP ARE EN ROUTE TO"
OPTIONS FOR NEXT WORD OR SYMBOL ARE:
<COUNTRY>,  <PORT>
end-of-list
        * LIFER cannot parse the question; "America" is not in the
        * vocabulary.

        * INLAND can answer questions about the grammar itself.
```

```
6_what is a <country>
PARSED!
"<COUNTRY>" may be any sequence of words following one of the patterns:
<COUNTRY> =>  THE <COUNTRY>
<COUNTRY> =>  U S
<COUNTRY> =>  U S S R
<COUNTRY> =>  U S S
<COUNTRY> =>  U S A
<COUNTRY> =>  U K
<COUNTRY> =>  SOVIET UNION
<COUNTRY> =>  UNITED STATES
<COUNTRY> =>  UNITED KINGDOM
<COUNTRY> =>  SOUTH AFRICA
<COUNTRY> =>  WEST GERMANY
<COUNTRY> =>  SAUDI ARABIA
<COUNTRY> =>  GREAT BRITAIN
<COUNTRY> =>  H M S
"<COUNTRY>" may be any member of the set {ANGOLA ANGOLAN ARABIA ARABIAN
ARGENTINA ARGENTINAN BRITAIN BRITISH CANADA CANADIAN DUTCH EGYPT EGYPTIAN
ENGLAND ENGLISH FOREIGN FRANCE FRENCH GERMAN GERMANY H.M.S. HMS ITALIAN ITALY
LIBERIA LIBERIAN NETHERLANDS NORWAY NORWEGIAN PORTUGAL PORTUGUESE RUSSIA RUSSIA
N
 SOVIET SPAIN SPANISH U.K. U.S. U.S.A. U.S.S. U.S.S.R. UK US USA USS USSR
VENEZUELA VENEZUELAN}
Finished

        * The user may add synonyms to the vocabulary.
7_define america like USA
PARSED!
FINISHED
8_redo 5
PARSED!
Parse time: .482 seconds
CCA1:FOR R1 IN NNMOVES WITH (DSC EQ 'US') FOR NSTDPORT1 , R2 IN NNSHIP WITH
CCA1:R2.UICVCN EQ R1.UICVCN BEGIN STRING1 = R2.NAM END;
*> Total bytes transferred: 256
Computation time for query: 1.488 seconds
Real time for query: 68.168 seconds
SHIP = KENNEDY JF, LOS ANGELES, BATON ROUGE, PHILADELPHIA, POGY, ASPRO, SUNFISH
, KAWISHIWI

9_how many of them are navy ships
               THEM => ((DSC EQ 'US'))
        * "Them" is always interpreted as any reference to a set of
        * ships in the previous query.
PARSED!
Parse time: .599 seconds
CCA1:FOR R1 IN NNMOVES WITH (DSC EQ 'US') AND
CCA1:((TYPE NE 'BULK') AND (TYPE NE 'TNKR')) FOR NSTDPORT1 , R2 IN NNSHIP WITH
CCA1:R2.UICVCN EQ R1.UICVCN BEGIN STRING1 = R2.NAM END;
*> Total bytes transferred: 256
Computation time for query: .623 seconds
Real time for query: 21.641 seconds
8 of them:
SHIP = KENNEDY JF, LOS ANGELES, BATON ROUGE, PHILADELPHIA, POGY, ASPRO, SUNFISH
, KAWISHIWI

10_give status kitty hawk
  Trying Ellipsis:  Ellipsis has failed.
THE PARSER DOES NOT EXPECT THE WORD "STATUS" TO FOLLOW
"GIVE"
OPTIONS FOR NEXT WORD OR SYMBOL ARE:
<RELATIVE.CLAUSE>,  <SHIP>,  <VALUE.SPEC>,  THE
end-of-list

        * Here is an example of the paraphrase feature of LIFER.  A
        * new pattern is defined by example.

11_define (give status kitty hawk )
 ...like (list the employment schedule, state of readiness, commanding
officer and position of kitty hawk)
 ...
PARSED!
Parse time: .705 seconds
        * The system answers the query as a side-effect of parsing
        * the paraphrase.
```

```
CCA1:OPEN %TOP.ACCAT.SAGALOWICZ.NSTDPORT3 WRITE;
CCA1:OPEN %TOP.ACCAT.NEMPSKD READ;
CCA1:OPEN %TOP.ACCAT.NREADINESS READ;
CCA1:OPEN %TOP.ACCAT.NTRACKHIST READ;
CCA1:CLOSE NNMOVES ;
CCA1:OPEN %TOP.ACCAT.NNUNIT READ;
CCA1:FOR R1 IN NNUNIT WITH (ANAME EQ 'KITTY HAWK') FOR R2 IN NTRACKHIST WITH
CCA1:R2.UICVCN EQ R1.UICVCN FOR R3 IN NREADINESS WITH R3.UIC EQ R2.UIC AND
CCA1:R3.UIC EQ R1.UIC FOR NSTDPORT3 , R4 IN NEMPSKD WITH R4.UIC EQ R3.UIC BEGIN
CCA1:STRING1 = R1.RANK STRING2 = R1.CONAM STRING3 = R2.PTP STRING4 = R2.PTD
CCA1:STRING5 = R3.READY STRING6 = R4.ETERM STRING7 = R4.EBEG STRING8 = R4.EEND
CCA1:END;
*> Total bytes transferred: 85
Computation time for query: 1.883 seconds
Real time for query: 145.701 seconds
(EMPLMNT SURVOPS EMPBEG 760103 EMPEND 760205 READY 2 RANK CAPT NAME SPRUANCE R
POSITION 3700N01700E DATE 7601171200)
        * The generalized pattern for the paraphrase is added to
        * the grammar.
LIFER.TOP.GRAMMAR  =>  GIVE STATUS <SHIP>
        * F0074 is the LISP function created as the response expression
        * for this pattern.
F0074 (GIVE STATUS <SHIP>)

        * The new pattern may now be used like any other.  All of the
        * LIFER features, including spelling correction, are applied
        * just as they are for the initial patterns.

12_give status u s cruisers in the mediteranean
                        spelling-> MEDITERRANEAN
PARSED!
Parse time: 2.362 seconds
CCA1:FOR R1 IN NNSHIP WITH (NAM EQ 'MEDITERRANEAN') FOR NSTDPORT1 , R2 IN
CCA1:NTRACKHIST WITH R2.UICVCN EQ R1.UICVCN BEGIN STRING1 = R2.PTP STRING2 =
CCA1:R2.PTD END;
*> Total bytes transferred: 27
CCA1:FOR R1 IN NNSHIP WITH (NAT EQ 'US') AND (TYPE1 EQ 'C') AND (TYPE2 NE 'V')
CCA1:FOR R2 IN NTRACKHIST WITH (PTP EQ '2131S00234E') AND R2.UICVCN EQ R1.UICVC
N
CCA1:FOR R3 IN NNUNIT WITH R3.UIC EQ R2.UIC AND R3.UIC EQ R1.UIC FOR R4 IN
CCA1:NREADINESS WITH R4.UIC EQ R3.UIC FOR NSTDPORT3 , R5 IN NEMPSKD WITH R5.UIC
CCA1:EQ R4.UIC BEGIN STRING1 = R2.PTP STRING2 = R2.PTD STRING3 = R3.RANK STRING
4
CCA1:= R3.CONAM STRING5 = R3.ANAME STRING6 = R4.READY STRING7 = R5.ETERM STRING
8
CCA1:= R5.EBEG STRING9 = R5.EEND END;
*> Total bytes transferred: 1840
Computation time for query: 2.794 seconds
Real time for query: 250.619 seconds
I have 16 answers; do you want to see them all? ...Yes
EMPLMNT EMPBEG EMPEND READY RANK NAME          POSITION    DATE       SHIP
SURVOPS 760103 760205 2     CAPT SPRUANCE R   3700N01700E 7601171200 KITTY HAWK
CARESC  760101 760601 1     CAPT MORRIS R     4000N00600E 7601171200 CALIFORNIA
        .
        . {other lines omitted}
        .
REPL    760115 760125 1     CAPT SHELL R      4000N00600E 7601171200 HASSAYAMPA
REPL    760109 760119 1     CAPT ARCO A       3700N01700E 7601171200 ASHTABULA

13_done
PARSED!
CCA1:^Z
*> Set parameters
*< Q    QUIT  [confirm with <CR>]
File closed 20-Jun-77 16:51:56
Thank you
@
```

END

FILMED

9-85

DTIC