The USL/DEMS NASA/RECON Working Paper Series contains a collection of reports representing results of activities being conducted by the Computer Science Department of the University of Southwestern Louisiana pursuant to the specifications of National Aeronautics and Space Administration Contract Number NASW-3846. The work on this contract is being performed jointly by the University of Southwestern Louisiana and Southern University.

For more information, contact:

Wayne D. Dominick
Editor
USL/DEMS NASA/RECON Working Paper Series
Computer Science Department
University of Southwestern Louisiana
P. O. Box 44330
Lafayette, Louisiana 70504
(318) 231-6308
NATURAL LANGUAGE QUERY SYSTEM DESIGN FOR INTERACTIVE INFORMATION STORAGE AND RETRIEVAL SYSTEMS: PRESENTATION VISUALS

I-Hsiung Liu

The University of Southwestern Louisiana Center for Advanced Computer Studies Lafayette, Louisiana

April 22, 1985
NATURAL LANGUAGE QUERY SYSTEM DESIGN

FOR

INTERACTIVE INFORMATION STORAGE

AND RETRIEVAL SYSTEMS

A Thesis

Presented to

The Graduate Faculty of

The University of Southwestern Louisiana

In Partial Fulfillment of the

Requirements for the Degree

Master of Science

I-Hsiung Liu

April 1985
OVERVIEW

* INTRODUCTION

* WHY STUDY CASUAL USER/SYSTEM INTERACTION

* RESEARCH OBJECTIVES AND METHODOLOGY

* SCOPE OF THE RESEARCH

* IMPORTANCE OF NATURAL LANGUAGE QUERY SYSTEMS DEVELOPMENT

* DYNAMICS OF INFORMATION STORAGE AND RETRIEVAL SYSTEMS

* FORMAL QUERY INTERFACE AND INFORMATION RETRIEVAL

* NATURAL LANGUAGE INTERFACE AND INFORMATION RETRIEVAL

* FORMAL QUERY APPROACH AND NATURAL LANGUAGE APPROACH
OVERVIEW (CONT'D)

* NATURAL LANGUAGE QUERY SYSTEMS DEVELOPMENT
* NATURAL LANGUAGE PROCESSING
* LANGUAGE CAPABILITIES OF NATURAL LANGUAGE QUERY SYSTEMS
* PHASES OF NATURAL LANGUAGE PROCESSING
* FRAMEWORK
* DESIGN METHODOLOGY
* INTERFACES OF NATURAL LANGUAGE QUERY SYSTEMS
* INFORMATION RESOURCES FOR NATURAL LANGUAGE PROCESSING
* FUNCTIONAL FEATURES OF THE SYSTEM
* CONCLUSIONS
WHY STUDY CASUAL USER/SYSTEM INTERACTION

- CASUAL USERS: THE POTENTIALLY LARGEST USER GROUP

- NATURE OF CASUAL USERS
  - IRREGULAR INTERACTION WITH THE SYSTEM
  - LACK KNOWLEDGE ABOUT COMPUTERS, PROGRAMMING, ETC.
  - NOT WILLING TO LEARN ANY ARTIFICIAL LANGUAGE

- PROBLEMS OF CONVENTIONAL MULTI-LEVEL QUERY SYSTEMS
  - HUMAN INTERFACES AS A ONE-WAY COMMUNICATION PROCESS
  - LANGUAGE INTERFACES FOR EXPERIENCED USERS
  - HUMAN INTERMEDIARY OFTEN STILL REQUIRED
OBJECTIVES AND METHODOLOGY

* OBJECTIVES

* PROPOSE A FRAMEWORK FOR EXPLORING NATURE, SCOPE, AND CONTENT OF THE EVOLVING TOPIC OF NATURE LANGUAGE QUERY SYSTEMS (NLQS)

* METHODOLOGY

* EXAMINE THE RATIONALE FOR NLQS DEVELOPMENT
* REVIEW CONCEPTS AND APPROACHES TO NLQS DEVELOPMENT
* PRESENT THE FRAMEWORK FOR NLQS DEVELOPMENT
SCOPE OF THE RESEARCH

* THE TARGET SYSTEM OF THE RESEARCH
* INFORMATION STORAGE AND RETRIEVAL SYSTEMS (IS&R SYSTEMS)
* REASONS FOR STUDYING IS&R SYSTEMS
  * WELL-DEFINED PROBLEM AREA
  * WELL-UNDERSTOOD SEMANTICS
  * WELL-STRUCTURED DATA ORGANIZATION
  * COLLECTIONS OF PUBLICATIONS
  * DATA OF NATURAL LANGUAGE TEXTUAL FORM
  * PRE-PROCESSED DATA
* THE PRIMARY FUNCTION OF IS&R SYSTEM
* INFORMATION RETRIEVAL
DYNAMICS OF IS&R SYSTEMS

* FOUR COMPONENTS OF MAN/COMPUTER COMMUNICATION PROCESS

* TASK OF INFORMATION RETRIEVAL

* USERS

* USER INTERFACES

* DATABASES
DYNAMICS OF IS&R SYSTEMS (CONT’D)

- LEVEL OF DIFFICULTY
  - COMPLEXITY
    - NUMBER OF STATES (NUMBER OF EVENTS OR WAYS OF ARRANGING THEM) IN A PARTICULAR APPLICATION
  - DETERMINISM
    - EXTENT TO WHICH THE OCCURRENCE OF EVENTS OR THEIR SEQUENCE CAN BE PREDICTED IN ADVANCE
DYNAMICS OF IS&R SYSTEMS (CONT’D)

* KNOWLEDGE REQUIRED

* DOMAIN-SPECIFIED KNOWLEDGE

* INFORMATION EXISTING IN THE DATABASE ENVIRONMENT E.G., WELL-STRUCTURED DATABASE STRUCTURES

* WORLD KNOWLEDGE

* INFORMATION EXISTING IN REAL WORLD ENVIRONMENT E.G., UNSTRUCTURED DATA OR LINGUISTIC DATA
DYNAMICS OF IS&R SYSTEMS (CONT’D)

* DATABASES

* MODELS OF CERTAIN REALITY

* LEVELS OF ABSTRACTIONS

  * ENSURE BOTH PHYSICAL AND LOGICAL DATA INDEPENDENCE

  * ALLOW RETRIEVED DATA CLOSER TO USER’S VIEW
DYNAMICS OF IS&R SYSTEMS (CONT’D)

* USERS

* END USERS

* CUSTOMERS OF COMPUTER SYSTEMS

* PROGRAMMERS

* BUILDERS OF APPLICATION PROGRAMS

* DATABASE ADMINISTRATORS

* MANAGERS OF INDIVIDUAL DATABASES
DYNAMICS OF IS&R SYSTEMS (CONT'D)

* USER INTERFACES

* CONCEPTS OF HIERARCHY OF USER LANGUAGES

* EACH INTERFACE IS DEFINED IN TERMS OF A LOWER INTERFACE

* EACH INTERFACE MAY SERVE AS THE BASIS FOR THE DEFINITION OF A HIGHER INTERFACE

* EXACTLY ONE INTERFACE SERVES AS THE ULTIMATE BASIS FOR ALL OTHER INTERFACES
DYNAMICS OF IS&R SYSTEMS (CONT’D)

* USER INTERFACES (CONT’D)

* ADVANTAGES OF THE CONCEPT

* IT IS ALWAYS POSSIBLE TO DEVELOP AN INTERFACE ON TOP OF THE CURRENT LANGUAGE HIERARCHY

* THE DEVELOPMENT OF A HIGHER INTERFACE REDUCES THE BURDEN ON END USERS

* THE DEVELOPMENT OF HIGHER INTERFACE IMPLIES THE REDUCTION OF SYNTACTIC RESTRICTION ON THE USE OF LANGUAGE FACILITIES

* TWO MAJOR TRENDS OF USER INTERFACES DEVELOPMENT
FORMAL QUERY INTERFACE AND INFORMATION RETRIEVAL

- TASK OF FORMAL QUERY DATABASE SEARCHING
- NARROW DOWN THE SEARCH SPACE
- MAJOR FUNCTIONS
  * EXIST / COUNT
- PRIMARY OPERATIONS -- FETCH OPERATIONS
- DEMANDS ON DATABASE DESIGN
  * OPTIMIZE THE SPEED OF FETCH OPERATIONS
    * INVERTED FILE STRUCTURES
    * SEARCH TERM CLASSIFICATIONS
- KNOWLEDGE REQUIRED
  * SYSTEM -- DOMAIN SPECIFIED KNOWLEDGE
  * USERS
    * FORMAL LANGUAGE SYNTAX
    * SEARCH TERMS AND INDEXES
    * BOOLEAN AND RELATIONAL OPERATIONS
N. L. INTERFACE AND INFORMATION RETRIEVAL

* PURPOSE OF NLQS DEVELOPMENT
  * COPE WITH THE NATURE OF CASUAL USERS
  * ALLOW USERS TO EXPRESS THEIR NEEDS BASED ON THEIR OWN PERCEPTIONS
  * ALLOW USERS PERFORM QUERIES BASED ON THEIR ABILITY TO DEAL WITH NATURAL LANGUAGE SYNTAX
  * ONE-STEP TRANSLATION PROCESS

* KEY ISSUES OF NATURAL LANGUAGE QUERY SYSTEM DEVELOPMENT
  * DEVELOP SYSTEM ABILITY TO PERFORM NATURAL LANGUAGE TRANSFORMATION
  * IMPACT OF THE DEVELOPMENT ON OTHER INFORMATION RETRIEVAL COMPONENTS IN THE SYSTEM
FORMAL QUERY APPROACH AND N. L. APPROACH

* SIMILARITY
  * BOTH ALLOW A USER TO SPECIFY "WHAT" HE WANTS THE MACHINE TO DO

* DIFFERENCES
  * USER ORIENTED VERSUS SYSTEM ORIENTED
  * USER REQUIREMENTS
  * NATURAL OF THE LANGUAGE FACILITY
  * LANGUAGE TRANSLATION PROCESS
ARGUMENTS AGAINST NATURAL LANGUAGE APPROACH

* AMBIGUOUS AND IMPRECISE
* TIME CONSUMING
* RECOGNITION OF THE IMPORTANCE OF NATURAL LANGUAGE APPROACH

"WHILE THE QUESTION OF THE EFFECTIVE INTERFACE BETWEEN MAN/MACHINE INTERACTION IS RAISED, THE ISSUES OF CONCERN ARE THE HUMAN FACTOR ISSUES."
NATURAL LANGUAGE PROCESSING

* ACTIVITIES OF NATURAL LANGUAGE PROCESSING
  * TRANSLATING NATURAL LANGUAGE STATEMENTS INTO THEIR FORMAL COUNTERPARTS
  * GENERATING NATURAL LANGUAGE RESPONSES TO USER'S REQUESTS

* INTELLIGENT USE OF SYSTEM KNOWLEDGE
  * TYPES OF SYSTEM KNOWLEDGE
  * USE OF SYSTEM KNOWLEDGE
NATURAL LANGUAGE PROCESSING (CONT'D)

* ISSUES RELEVANT TO NATURAL LANGUAGE PROCESSING
  * LANGUAGE CAPABILITIES
  * PHASES OF NATURAL LANGUAGE PROCESSING
  * OPERATIONS REQUIRED FOR EACH PHASE
  * KNOWLEDGE APPLIED IN EACH PHASE
LANGUAGE CAPABILITIES OF NLQS

- DIFFICULTIES OF IDENTIFYING LANGUAGE CAPABILITIES
- COMPLEXITY OF NATURAL LANGUAGE
- UNPREDICTABILITY OF HUMAN ERRORS
- LANGUAGE CAPABILITIES EXHIBITED IN MOST SYSTEMS
  - ANSWER DIRECT QUESTIONS
  - HANDLE SIMPLE USE OF PRONOUNS AND ELLIPSIS
  - ANALYZE NULL ANSWERS
  - RESTATEMENT OF USER INPUTS
  - CORRECT TYPOGRAPHIC OR SPELLING ERRORS
PHASES OF NATURAL LANGUAGE PROCESSING

* SYNTACTIC ANALYSIS

* LEXICAL ANALYSIS -- CLEAN UP THE USER INPUT

* WORD BY WORD TRANSFORMATION

* LEXICON

* PARSING -- GENERATE TREE STRUCTURE OF INPUT

* TEMPLATE MATCHING PARSER

* PHRASE STRUCTURE GRAMMAR PARSER

* TRANSFORMATIONAL GRAMMAR PARSER

* SEMANTIC GRAMMAR PARSER
PHASES OF NATURAL LANGUAGE PROCESSING (CONT'D)

* SEMANTIC ANALYSIS

* GENERATE INTERNAL TARGET REPRESENTATIONS FROM NATURAL LANGUAGE INPUTS

* TYPES OF INTERNAL TARGET REPRESENTATIONS
  * WEIGHTED VECTORS
  * FIRST-ORDERED PREDICATES
  * SEMANTIC NETWORKS
  * CASE FRAMES
PHASES OF NATURAL LANGUAGE PROCESSING (CONT’D)

* EXECUTION PHASE

* PERFORM DATABASE SEARCHING

* APPROACHES TO EXECUTION
  * DIRECTLY USING INTERNAL TARGET REPRESENTATION
  * USING BOOLEAN CONDITIONS
  * USING SYSTEM-DEFINED FORMAL QUERY REPRESENTATIONS
PHASES OF NATURAL LANGUAGE PROCESSING (CONT’D)

* RESPONSE GENERATION
  * PROVIDE PRECISE NATURAL LANGUAGE RESPONSE
  * TYPES OF RESPONSES
    * DETECTED/CORRECTED ERRORS
    * RESTATEMENT OF LOGICALLY COMPLETED QUERY
    * DIRECT AND CORRECT ANSWERS TO THE QUERY
    * COOPERATIVE RESPONSES
FRAMEWORK

* GOALS OF THE FRAMEWORK
* DESCRIBE DESIGN METHODOLOGY
* STATE MAJOR DESIGN PROBLEMS
* DESCRIBE INTERFACES OF THE SYSTEM
* DESCRIBE FLOW OF CONTROL
* IDENTIFY SYSTEM COMPONENTS AND THEIR FUNCTIONS
* DESCRIBE RELATIONSHIPS BETWEEN SYSTEM COMPONENTS
DESIGN METHODOLOGY

* PROBLEMS OF TRADITIONAL APPROACH
  * UNABLE TO PROVIDE COOPERATIVE ANSWERS
  * UNABLE TO "UNDERSTAND" THE USER'S INPUTS
  * RESTRICTED NATURAL LANGUAGE INPUTS

* BASIC CONSIDERATIONS OF THE DESIGN
  * SYSTEM'S ABILITY TO "UNDERSTAND" NATURAL LANGUAGE
  * SYSTEM'S REQUIRED LINGUISTIC KNOWLEDGE
    * GRAMMAR RULES
    * SEMANTIC RULES
    * INTERNAL REPRESENTATIONS OF NATURAL LANGUAGE
DESIGN METHODOLOGY (CONT’D)

- * DESIGN APPROACH
  - * EXPERIMENTAL APPROACH
    * RESTRICTED DOMAIN KNOWLEDGE
    * RESTRICTED LANGUAGE CAPABILITIES
  * MAJOR DESIGN PROBLEMS
    * REDUNDANT WORDS OR PHRASES
    * AMBIGUITY AND VAGUENESS
    * ERROR DETECTION AND CORRECTION
INTERFACES OF NLQS

* NATURAL LANGUAGE INTERFACE

* INTERFACE OF CASUAL USER/SYSTEM INTERACTION

* MAJOR CONCERNS

* SYNTACTIC ANALYSIS OF NATURAL LANGUAGE INPUTS

* INTERPRETATION OF NATURAL LANGUAGE INPUTS

* NATURAL LANGUAGE RESPONSE GENERATION

* USER/SYSTEM DIALOGUE CONTROL
INTERFACES OF NLQS (CONT'D)

* FORMAL LANGUAGE INTERFACE
* INTERMEDIATE INTERFACE BETWEEN NATURAL LANGUAGE INTERFACE AND DATABASE INTERFACE
* MAJOR CONCERNS
  * TARGET REPRESENTATION AND FORMAL QUERY TRANSFORMATION
  * PORTABILITY
* DATABASE INTERFACE
  * EXECUTION LEVEL
INFORMATION RESOURCES FOR NL PROCESSING

* DOMAIN SPECIFIED KNOWLEDGE
  * INFORMATION ABOUT PART OF A DOCUMENT FILE
  * SOURCES OF THE KNOWLEDGE
    * BIBLIOGRAPHIC DATABASES
    * INVERTED FILES AND THESAURUS
* WORLD KNOWLEDGE
  * LINGUISTIC KNOWLEDGE
    * INFORMATION ABOUT A DOCUMENT FILE
    * SOURCE OF THE KNOWLEDGE
    * SEMANTIC NETWORKS
FUNCTIONAL FEATURES OF THE SYSTEM

* PARSER
* INTERPRETER
* FORMAL QUERY GENERATOR
* NATURAL LANGUAGE RESPONSE GENERATORS
* ERROR GENERATOR
* RESTATEMENT GENERATOR
* REPLY GENERATOR
FUNCTIONAL FEATURES OF THE SYSTEM (CONT’D)

* PARSER

* LEXICAL ANALYZER

* DELETE NOISE WORDS

* REPLACE INPUT WORDS BY THEIR LEXICON CLASS ENTRIES

* SUBSTITUTE ROOTS AND INFLECTION MARKERS

* PHRASE CONSTRUCTION ROUTINE

* SPELLING CORRECTION AND ENTRY ADDITION ROUTINE
FUNCTIONAL FEATURES OF THE SYSTEM (CONT’D)

* PARSER (CONT’D)
  * ATN PARSER
    * NON-DETERMINISTIC ANALYSIS
    * RESOLVE LEXICAL AMBIGUITY
    * QUALIFIER HANDLING
    * ELLIPSIS AND PRONOUN REFERENCE HANDLING
  * FAILURE OF PARSING
FUNCTIONAL FEATURES OF THE SYSTEM (CONT'D)

* INTERPRETER
  * GENERATE A SUBSET OF SEMANTIC NETWORK FOR USER INPUTS
* CASE FRAMES
  * REPRESENTING A SUBSET OF SEMANTIC NETWORK IDENTIFIED BY THE VERB OR CHARACTERISTICS-
  * CONSISTING OF A LIST OF INTERRELATED NP’S
* SEMANTIC AMBIGUITY HANDLING
* FAILURE OF SEMANTIC ANALYSIS
FUNCTIONAL FEATURES OF THE SYSTEM (CONT’D)

* FORMAL QUERY GENERATOR
  * TRANSLATE FILLED-IN CASE FRAMES INTO ONE OR MORE FORMAL QUERY EXPRESSIONS
  * DETERMINE THE TYPES OF THE USER REQUEST
  * SELECT SPECIFIC SEARCH COMMAND WORD(S)
  * SELECT SEARCH TERMS

* EXAMPLE

CONTAIN: DESTINATION

  = TEXT TERM(SUBJ) = "COMPUTER"

OBJECT - DOC REF CODE(OBJ) = ?

FORMAL QUERY:

SELECT TEXT TERM EQ COMPUTER
FUNCTIONAL FEATURES OF THE SYSTEM (CONT’D)

* NATURAL LANGUAGE RESPONSE GENERATORS
  * GENERATE APPROPRIATE FEEDBACK TO THE USER
  * ERROR GENERATOR
    * SYNTACTIC ERRORS
    * SEMANTIC ERRORS
  * RESTATEMENT GENERATOR
    * TO OBTAIN THE USER’S APPROVAL
    * RESTATEMENT CONSTRUCTION
      -- CASE FRAMES
  * REPLY GENERATOR
  * OUTPUT SEARCH RESULTS
CONCLUSION

* ACHIEVEMENTS OF THE FRAMEWORK
* PROVIDE A SOLID THEORETICAL FOUNDATION FOR IS&R NLQS DEVELOPMENT
* CONSTRUCT A SOUND IS&R NLQS ARCHITECTURE

* THE THEORETICAL FOUNDATION
* DYNAMICS OF INFORMATION RETRIEVAL
* LEVELS OF ABSTRACTIONS
* HIERARCHY OF USER LANGUAGES
* NATURAL LANGUAGE PROCESSING
* EXPERIMENTAL APPROACH TO SYSTEM DEVELOPMENT
CONCLUSION (CONT’D)

* THE ARCHITECTURE

* MAJOR ASSUMPTIONS

* NLQS DEVELOPMENT IS TO BUILD A SYSTEM WHICH CAN SIMULATE HUMAN BEHAVIOR

* NLQS DEVELOPMENT WOULD AFFECT THE DYNAMIC RELATIONSHIPS AMONG SYSTEM COMPONENTS

* INTERFACES

* PORTABILITY

* EXTENDABILITY

* FUNCTIONAL MODULES

* FUNCTIONAL DECOMPOSITION

* BLACK BOX APPROACH
* INTERPRETER

WRITE: DESTINATION

= AUTHOR (OBJ)

OBJECT = AUTHOR (OBJ)

WRITE ---> AUTHOR ---> "MARTIN, T."

OBJ

+---> ARTICLE

"INFORMATION RETRIEVAL"
<table>
<thead>
<tr>
<th>INDEX TERM</th>
<th>DOC REF CODE</th>
<th># DOC'S</th>
</tr>
</thead>
<tbody>
<tr>
<td>COMPUTER</td>
<td>3, 5, 8, 9</td>
<td>4</td>
</tr>
<tr>
<td>INFORMATION</td>
<td>1, 2, 4, 7, 9</td>
<td>5</td>
</tr>
<tr>
<td>:</td>
<td>:</td>
<td>:</td>
</tr>
<tr>
<td>:</td>
<td>:</td>
<td>:</td>
</tr>
</tbody>
</table>
+----------+
| USER     |
+----------+
    |
| natural language |
| query       |
+-------------------+
| NATURAL LANGUAGE |
| INTERFACE       |
+-------------------+
    |
| internal representation |
| (semantic nets)    |
+-------------------+
| FORMAL LANGUAGE  |
| INTERFACE       |
+-------------------+
    |
| formal query     |
+-------------------+
| DATA BASE       |
| INTERFACE       |
+-------------------+
    |
| DATA BASE       |
Fact

REALITY

Information Structure

USER'S LOGICAL LEVEL

Information Structure Description Language (ISDL)

(logical data independence)

Data Structure

SYSTEM'S LOGICAL LEVEL

Data Structure Description Language (DDL)

(physical data independence)

Storage Structure

PHYSICAL STORAGE LEVEL

Storage Structure Description Language
<table>
<thead>
<tr>
<th>METHODOLOGY</th>
<th>CHARACTERISTICS</th>
</tr>
</thead>
<tbody>
<tr>
<td>FACT RETRIEVAL</td>
<td>* PRE-PROCESSED DATA</td>
</tr>
<tr>
<td></td>
<td>* EXTENSIVE USE OF FETCH OPERATION</td>
</tr>
<tr>
<td></td>
<td>* RESULTS DIRECTLY OBTAINED FROM DATABASES</td>
</tr>
<tr>
<td></td>
<td>* SIMPLE REQUEST-RESPONSE SEQUENCE</td>
</tr>
<tr>
<td>STATISTICAL INFERENCE</td>
<td>* WELL-STRUCTURED RAW DATA</td>
</tr>
<tr>
<td></td>
<td>* EXTENSIVE USE OF GET-NEXT OPERATIONS</td>
</tr>
<tr>
<td></td>
<td>* RESULTS OBTAINED FROM DATABASE REFERENCING</td>
</tr>
<tr>
<td></td>
<td>* DATA REDUCTION PROCESS</td>
</tr>
<tr>
<td>DEDUCTIVE INFERENCE</td>
<td>* SEMI-STRUCTURED DATA AND RULES</td>
</tr>
<tr>
<td></td>
<td>* SYSTEM MUST ANSWER WHY-QUESTIONS</td>
</tr>
<tr>
<td></td>
<td>* RESULTS DEPENDENT ON THE SEARCH FOR</td>
</tr>
<tr>
<td></td>
<td>POSSIBLE RELATIONS BETWEEN DATA</td>
</tr>
<tr>
<td>Task</td>
<td>Documental Retrieval</td>
</tr>
<tr>
<td>--------------------</td>
<td>----------------------</td>
</tr>
<tr>
<td>Retrieval Methodology</td>
<td>Fact Retrieval</td>
</tr>
<tr>
<td>Retrieval</td>
<td>Statistical Inference</td>
</tr>
<tr>
<td>Knowledge Required</td>
<td>Domain Specified Knowledge</td>
</tr>
<tr>
<td>Level of Difficulty</td>
<td>Low</td>
</tr>
</tbody>
</table>

This report represents one of the 72 attachment reports to the University of Southwestern Louisiana's Final Report on NASA Grant NGT-19-010-900. Accordingly, appropriate care should be taken in using this report out of the context of the full Final Report.