NATURAL LANGUAGE FOR DATABASE QUERIES:

A LABORATORY STUDY

Yannis Vassiliou, Matthias Jarke, Edward A. Stohr,

Jon A. Turner, and Norman H. White

July 1983

Center for Research on Information Systems
Computer Applications and Information Systems Area
Graduate School of Business Administration
New York University

Working Paper Series

CRIS #55

GBA #83-74(CR)

Published in MIS Quarterly, Vol.7, No.4 December (1983) pp.47-61

This work was carried out as part of a joint study being conducted with the IBM Information Systems Group in White Plains, New York.

ABSTRACT

Are natural language systems for database queries meeting their goals? And, are these goals appropriate? The recently completed Advanced Language Project at New York University combined a field experiment with two laboratory studies to examine these issues by comparing performance between subjects using the formal database language SQL and subjects using the prototype natural language system, USL. This paper describes the design and results of the larger laboratory experiment. The results presented offer some promise for the usability of natural language under certain conditions.

Natural language (for example, English) appears to be the most controversial among the language interfaces that have been proposed for direct interaction with databases. Due to the task characteristics of database querying, natural language query systems have different structure and goals than other computer natural language systems, such as systems for text generation. Are these query systems meeting their design goals? More importantly, are these the "appropriate" goals? These seem to be the major questions for which no conclusive answers have yet been given.

A recently completed study at New York University constitutes a step toward resolving some of the issues pertaining to the use of natural language for database queries. The overall approach involves a combination of exploratory field evaluations with controlled laboratory studies. After a brief survey of natural language query systems and issues, this paper describes in detail a laboratory study which was conducted as part of the project.

BACKGROUND ON NATURAL LANGUAGE SYSTEMS

The term "natural language system" has been used to refer to many computer systems in drastically different application domains, e.g. foreign language translation, text generation, computer programming, conversational problem solving, and question-answering. Even though all these systems have a common characteristic, namely a natural language (German, English, etc.) interface, they have different goals, and therefore exhibit unique properties. For instance, a system having the goal of generating poetry will be inappropriate for

conversational problem solving. As a consequence, it is important to isolate the issues in each category of natural language systems and to explore the usability of natural language in the limited domain.

Natural language (NL) systems for querying a database have shown technical feasibility and promise of practical use, as evidenced by the large number of experimental systems [2, 7, 12, 19, 32, 33], and the commercial availability of at least one such system [1]. Yet, there is no scientific evidence to permit conclusive statements as to the usability of natural language for database inquiries.

Even among NL systems for querying a database, succinct philosophical differences exist. These are discussed in the context of the NL system under study.

The Design Goals of Natural Language Systems for Databases

The system used for the experiment is USL (User Specialty Languages), a prototype natural language query system [12, 18]. USL's aims can be summarized as: 1) economically allowing users to issue questions (queries) to a database in a natural language (e.g., German, English, Spanish), and 2) to quickly receive well-formatted meaningful responses. The design goals and structure of USL, shared also by other NL systems (e.g. INTELLECT [1]), are described below.

Type of System - The rationale for providing a natural language interface is to give users direct access to databases. Frequent routine queries can often be incorporated into simpler, menu-driven

systems; it is the applications with non-standard, ad-hoc queries to which USL is directed. In this respect, the user interaction with USL is similar to that of formal query languages required by most database systems (e.g. the language SQL [5]).

The developers of USL put the onus of understanding the natural language almost entirely on the system itself. They aimed to avoid the clarification dialogue approach to language understanding, where each query the user poses is followed by an often-lengthy dialogue between the system and the user in order for the system to interpret the original query [3, 33]. Figure 1 presents an example of the interaction between USL and its users. The same dialogue using SQL is presented in Figure 2.

```
...
USER:   how many alumni have no donations?

SYSTEM:
          3679

USER:   list all the alumni who live in detroit

SYSTEM:   | LASTNAME  | FIRSTNAME  | CITY    |
          ------------------------------------
          | jones     | douglas    | detroit |
          | ...       | ...        | ...     |

          ...
```

Figure 1: USL Session

```
          ...
  USER:   select count(unique id)
          from   donors
          where  donors.srccode = 'al' and
                 donors.id not in
                        (select unique id
                         from   giftsummary);

SYSTEM:   ----------
             3679

  USER:   select lastname, firstname, city
          from   donors
          where  donors.srccode = 'al' and
                 donors.city = 'detroit';

SYSTEM:   | LASTNAME  | FIRSTNAME  | CITY    |
          -------------------------------------
          | jones     | douglas    | detroit |
          | ...       | ...        | ...     |

          ...
```

Figure 2: SQL Session

Type of Users - The intended users of USL are application specialists, including analysts, clerical workers, planners, and management. The system was planned for both frequent and infrequent users. Accordingly, USL is designed so that users would not be required to have any programming or database skills or extensive linguistic knowledge. The only demand USL purports to make on the user is sufficient familiarity with the application to be able to compose meaningful queries.

Type of Use - The developers' aspiration was to develop a system structure that would enable USL to be transferred to new applications and to other natural languages (e.g. Spanish) quickly and economically. This goal distinguishes USL from the Special Purpose Language and Data Retrieval Systems [35], which are highly tailored to the application context and therefore require large economic and time commitments for each implementation. At the other extreme, the developers of USL also rejected the prohibitive nature of a system encompassing the entire English language. Instead, the developers tried to strike a balance between these polar cases in attempting to design a strong core system for analyzing English queries, while requiring some application-specific vocabulary to be added upon installation of the system. In this way, the User Specialty Languages system is intended to economically meet the modes of expression of each particular user group.

Generality and portability are primarily achieved by separating the linguistic component from the database system. Only structural database information is required for the language analysis. Natural language expressions are then mapped directly to high-level database language expressions. USL sits on top of a generalized database management system, and is translated to the formal language SQL [5]. It should be noted that USL does not have a general purpose deductive mechanism that makes inferences from an artificial intelligence-based knowledge representation. The utility of such a mechanism is traded-off for speed of execution, transportability to many applications, and advanced facilities offered by the database management system (e.g. calculations).

USL's structure and goals are shared by other general purpose database query systems using natural language, e.g. TEAM [7], IRUS [2], and INTELLECT [1]. These goals have been criticized by several researchers [23] as severely limiting the principle of natural language use, and in the long run being impractical. Tennant [28] writes:

> "...without extending conceptual coverage beyond the limits of the database contents, a natural language question answerer can do little more than a formal query language. What's worse is that the natural language version would be more expensive to run..."

Therefore, the issue of NL for database querying raises two important research questions: (1) Are the goals set forth by NL systems the "right" ones? and, (2) how well are these goals met by such systems?

Experimental Studies for Natural Language Usability

Most experimental studies with NL systems have focused on the question of whether the system under study meets its goals.

For instance, Tennant [28] reports that in a laboratory study with novice-application specialists using PLANES [33], 275 queries out of 402 were understood correctly by the system. Of the 117 errors only 40 were attributed to inadequacies of the PLANES system.

Damerau [4] described the results of running the Transformational Question Answering System (TQA, formerly REQUEST) in a city government planning department. Of 788 queries posed to the system over a twelve month period, 513 or 65 percent were successfully completed. No

information is given on how subjects were trained, what assistance they were given during the experiment or how queries were scored.

The LADDER system was evaluated by Miller et al [16] as a database query language and it was shown that users were able to use the system with some facility after an hour and a half instruction. The emphasis was placed on skill acquisition (learning) as in many experiments with formal database query languages [20, 21, 22, 29, 34].

High success rates have been reported in field tests with NL systems. For instance, Harris [8] reports 80-90% successful queries with the ROBOT system (precursor of INTELLECT). Similarly, Krause [11] and Lehmann [13] report over 80% success with the German version of USL.

Egly and Loebner [6] performed an analysis of four protocols of subjects using REL [30]. They found that subjects were able to use their knowledge of natural English to discover how the features of the engineered REL relate to database access; how the lexicon pertains to the retrieval mechanisms, which grammatical constructs are permitted, and which constructs are semantically equivalent paraphrases.

Some laboratory studies did not consider any specific system and attempted to investigate the usability of natural language for database queries. For instance, Shneiderman [25] briefly trained subjects in SQL and then tested them in an experiment to determine whether they asked more valid queries in English than in SQL. He found no significant difference in the number of valid queries asked, but did find an order effect with the English-SQL group having more

errors than the SQL-English group. Also, Small and Weldon [26] reported on a laboratory study where novices were tested on a simulated processor. Productivity of natural language versus SQL was the major research question, and a superiority of SQL in query formulation time was observed.

Malhotra [14] conducted a simulation study to assess user requirements for NL communication with computers. One of his conclusions was that "any system that purports to allow convenient conversational interaction in English must be able to deal with pronoun and anaphoric reference, and ellipses." (p.168) Malhotra emphasizes the need for the system to possess domain-specific knowledge so that it can respond intelligently and flexibly to ambiguous user requests. He also states that making the system natural to use should include protecting the naive user from system errors and their associated cryptic messages such as "ERROR 1273 ILLEGAL REFERENCE FROM 1623".

While these studies provide some useful insights about natural language much remains to be investigated. Tennant [28] is critical of the lack of exploratory studies:

> The lack of evaluation of natural language processing research leave several critical questions about the work unanswered. Readers are unsure what concepts are included in the system, what accomodations have been made for language variations between users, the restrictions on the discourse domain or database, the restrictions on data manipulation capabilities, and the restrictions on inferencing capabilities. There is usually no information about the match between facilities included in the system and the actual needs of the users. In addition there is little information on what kind of performance would be required of a natural language processor to allow users to carry out tasks at various levels of complexity (p. 3).

A Research Project Combining Laboratory and Field Evaluations

We have argued that there are two major questions for the usability on natural language as a database query language. First, are natural language systems setting the appropriate goals in attempting to meet the user needs?, and second, do they meet these goals?

A negative answer to the second question, as is usually the case with prototype systems, makes the determination of an answer for the first question very difficult. Even though field tests offer more promise than laboratory studies in assessing the usability of natural language systems, they are often hampered by implementation limitations, and of course, by the lack of a controlled environment.

We see the combination of exploratory field evaluations with laboratory studies as a strong research strategy to investigate the usability of natural language for database queries. Exploratory studies in real work settings offer the most likely means of identifying critical issues for more detailed study in laboratory experiments. This was the approach taken for the Advanced Language Project (ALP) where a field test was conducted, together with two laboratory experiments [27]. This paper describes the results of the second laboratory experiment.

Rather than attempt to evaluate a natural language application in the absolute, it was decided to compare the performance of subjects using natural language to the performance of another group of subjects using a reference artificial language with the same application.

Since USL maps natural language queries to SQL for a database access, and SQL has been extensively studied [21, 34], it was decided to use SQL as the reference (comparison) language.

The application domain selected for ALP was a Question-Answering system about Alumni of the Graduate School of Business Administration at New York University. The system maintains demographic and donation history data of school alumni, foundations, other organizations, and individuals. The school has over 40,000 graduates as well as over 5,000 non-graduates who have given to the school over the past 20 years. Eight intermediaries for the principal users of this application (Deans and development officers), were the subjects for the exploratory study.

THE LABORATORY EXPERIMENT

Preliminary results from the Advanced Language Project [27, 31] indicated several issues that needed further investigation and could be better tackled in a laboratory setting. In particular, three major issues were identified:

First, due to the large size of the field study, it was not possible to make a detailed evaluation of the conceptual methodologies employed by the users and of the word usage in requests. Word usage is very important for the design of a language system such as USL and for the development of USL applications. USL provides a set of application independent words as a core. It is the responsibility of application developers to add the words that pertain to a particular

application. For these two phases of creating the lexicon some guidance is needed. Also, the question often arises: can users be restricted to this lexicon without any behavioral difficulties?

Second, the generally hostile operator environment of the field experiment undoubtedly introduced a large number of errors. Line problems, printing delays and long system delays negatively biased the language evaluation. Such bias is not present in a pencil-and-paper laboratory experiment.

Third, in the previous laboratory study and field test, it was observed that "training" in USL was necessary (USL is sufficiently demanding in its restrictions). This new study presented the opportunity to test our training methodology.

Research Questions and Hypotheses

The laboratory study explores the following hypotheses.

H1: There will be no difference in performance between subjects using USL and those using SQL.

A paper and pencil test represents an idealized situation. The formality of SQL offsets the potential confusion created by having to learn arbitrary restrictions in USL. Also, all negative effect factors for performance (bad interface, no constructive feedback, etc.) in a field study are eliminated in a pencil and paper test. These factors affect USL more than SQL.

H2: The query lengths for subjects using SQL will be greater than the ones for subjects using USL.

The SQL user is required to stay within the framework imposed by the syntax of the language; all needed keywords have to be referred to, and often precise disambiguation of attribute names (e.g. DONORS.ID as opposed to simply ID) is necessary. On the other hand, in English the user can use sentence-fragments instead of complete sentences. For instance, the imperative verb may be omitted, adjectives may replace qualifications, etc. The laboratory experiment permitted testing of this hypothesis for English, rather than strictly for USL which does not accept all fragments. The subjects' solutions need not be accepted by the USL system.

H3: SQL subjects require more query formulation time than USL subjects.

It should be expected that the direct adherence to SQL syntax, the verboseness of SQL, and its procedurality will result in higher query formulation times for SQL than for USL.

H4: Training in USL (in addition to application training) is necessary.

Training in USL consists of learning language and system restrictions. If no such training is given, subjects may use the language procedurally, and may employ language constructs not supported by the USL system (e.g. modality, passive voice).

In addition to the above testable hypotheses, the laboratory study allows for the investigation of the following research questions:

RQ1: Can a restricted vocabulary be enforced for the use of English, without behavioral difficulties?

This question has also been explored in [17] and [10]. In essence, it refers to the basics of USL's philosophy; the possibility of defining a "manageable" vocabulary. The type of words (grammar categories) used by subjects indicate where emphasis should be placed in language and application design.

RQ2: Do subjects have similar conceptual problem solving frameworks (within the same language type)?

This question could be partially answered by the number of words used per question and per subject, as well as, by investigating the commonality of word usage and strategies employed by subjects in answering a question.

Description of the Study

A group of 61 students with little or no prior computing experience were selected as paid subjects. These type of users have been referred to as "novice-casual" [32]; they have little knowledge of either programming concepts or of the application domain. The subjects were divided in three sub-groups:

G1: USL with application training (10)

G2: USL with application and language training (34)

G3: SQL with application and language training (17)

The number of USL subjects was larger because a continuation of the study was planned in which two groups of trained USL subjects were required. The assignment of subjects to groups was random with approximately even number of men and women, and mean years of age and

work experience. The groups were trained for two hours in the application domain (alumni donations). In addition, groups two and three were trained in their respective languages for three and one half hours. Subjects in group one were given a ten minute introduction to the interaction philosophy of USL (i.e. the fact that it is a question-answering system). This group was only used to test hypothesis H4.

All treatment groups were given the same paper and pencil test consisting of fifteen questions. Subjects were required to write the queries that were needed to answer the questions in their assigned language. Subjects were also asked to indicate on a five point scale the extent of their understanding of the question (clarity), how certain they were of a solution strategy, and how complex they believed the question to be. The exams were graded by two examiners.

Method - Each question in the exam was designed with no bias toward USL or SQL. Questions described problem situations with which the subjects had become familiar during their training. Subjects were asked to express a query (or a series of queries) to answer the question. For example:

> Q6.- A list of alumni in the state of California has been requested. The request applies to those alumni whose last name starts with an "S". Obtain such a list containing last names and first names.

The problem situation has three parts. First, the context is given. Second, some clues for the query are presented. The actual action to be taken is described in the third part. Since the information to

compose the query is scattered, the answer is not given away to the USL subjects. Correct answers in USL (English) and in SQL for the above request are:

Q6.- (USL). "What are the last names and first names of all California Alumni whose last name is like S% ?"

```
Q6.- (SQL). "Select lastname, firstname
             From donors
             Where srccode = 'al' and state = 'ca' and
                   lastname like 's%';
```

Questions differed in their degrees of difficulty and were placed in a constrained random order with an easy question first and a hard question last. Care was taken to include requests covering a wide range of language constructs. Written instructions and hints were given, together with reference material.

Training in SQL was similar to the approach adopted in [34] and [20]. It basically consisted of a number of examples after the syntax was learned. Training in USL mainly consisted of examples to work around language restrictions. Little emphasis was placed on the enumeration of the capabilities of USL (What can you do in English?). Rather, the emphasis was on presenting the system's basic characteristics (e.g., interaction, lack of intelligence), and the major language constructs not supported (e.g., sentence fragments, modality).

Measures in grading of exams - Five different measures were used (see below for details):

1. Correctness (scale: 1-10)
2. Welty-Correctness
3. Grammatical Correctness (scale: 1-5) USL-only
4. English Naturalness (scale: 1-5) USL-only
5. Time and Subjective Measures

CORRECTNESS (1-completely incorrect, 10-correct). - A measure of how close to a running USL/SQL query the subject's solution is.

WELTY-CORRECTNESS. - This measure differs from the "Correctness" measure in that it also attempts to specify a cause for the solution's errors. Thus solutions can be grouped in different categories. Another obvious advantage is the compatibility of our results with those of Welty's experiment [34]. If queries are coded using the first four codes, they are called "essentially correct". The codes (adjusted to our experiment) are:

'PR' - The solution is completely correct
'ML' - The solution is basically correct. Any small error would have been detected and possibly corrected by a good system, e.g. misspelling.
'MO' - The solution is again basically correct. It may contain a small error in data specifications, e.g. Bston instead of Boston. In this case, the output would have been null.
'MS' - The solution contains a minor substance error. Query output would have been incorrect, but the error is possibly due to the statement of the problem, or a language inadequacy.
'CO' - Correctable. The solution is wrong but a good system would have helped the user correct any syntactic errors.
'XS' - Major Substance Error. The solution is not for the request at hand but for a different one.
'XF' - Major Language Error. The solution does not follow the rules of the language used.
'IN' - Incomplete Solution.
'UN' - No attempt was made for a solution

GRAMMATICAL CORRECTNESS - This is a subjective measure of the grammaticality of the subject's solution. A value '5' indicates a 'correct' English query, while the value '1' indicates a completely ungrammatical sentence, even though it might be unambiguous in human communication.

ENGLISH NATURALNESS. - This is a measure relating to the difference between 'competence' and 'performance' in the use of English. There are expressions that use the English grammar rules to the letter (competence), but may be awkward or too verbose, and therefore are not natural (performance). Syntactically correct but otherwise unnatural solutions were given a low grade on a scale of 1 to 5.

TIMING AND SUBJECTIVE MEASURES. - Subjects were asked to record the time taken for each question, as well as their perception of request clarity, complexity, and their confidence of a solution strategy.

Analysis of Word Usage - In addition to looking at the subjects' solutions at the conceptual methodology and correctness levels, the most elemental aspects of the solutions were considered: individual words. General characteristics of the words used were explored: e.g., total number of words, total number of unique words, syntactic categories of words, frequency of word use per syntactic category, commonality of word usage, etc. All these are important for application development in USL. For the analysis of word usage, the solutions of all 17 SQL subjects were used, and compared with the solutions of 17 USL subjects randomly selected from the group of subjects trained in the application and the language.

Language Performance Results

Hypothesis H1: No significant difference in test scores was found between treatments (see Table 1). When the Welty category scale was used no significant difference between the test scores of treatment groups was found either (see Table 1). The two scoring methods are highly correlated (r=.864, p=.000, n=1048) on a question-by-question basis. There were 44.6% and 53.3% "essentially correct" queries (queries coded with Welty-codes 'PR', 'ML', 'MO', and 'MS') in USL, SQL respectively. In addition, there were few significant differences in performance for individual questions and overall they favored no language in particular (see Table 2). These results give support to the hypothesis.

CORRECTNESS	Mean	S. D.
USL	6.89	2.31
SQL	7.14	2.27
WELTY-SCALE	**Mean**	**S. D.**
USL	5.60	2.39
SQL	5.89	2.05

Table 1: Overall Performance Scores

Hypothesis H2: Verboseness was not a characteristic of English usage. There was an average of 21.2 words used for USL requests as compared to an average 33.8 of words used for SQL requests. These results support the hypothesis.

qu. no.	USL score mean	s.d.	SQL score mean	s.d.	t	p	USL better(+), SQL better(-)
1	8.9	1.3	9.2	1.2	-0.70	.487	
2	7.8	1.8	5.8	2.3	3.03	.006	+
3	7.6	2.2	6.7	2.2	1.32	.197	
4	8.0	1.9	6.5	2.3	2.35	.026	+
5	8.9	1.6	8.2	1.5	1.52	.139	
6	7.0	1.5	8.3	2.0	-2.35	.027	-
7	6.3	1.9	6.6	2.2	-0.55	.583	
8	5.8	1.8	5.9	2.3	-0.26	.801	
9	5.3	2.3	5.8	2.7	-0.64	.530	
10	7.1	2.4	5.7	2.3	2.02	.051	
11	6.5	2.7	8.1	2.2	-2.30	.027	-
12	7.7	2.2	8.8	1.7	-1.95	.058	
13	6.3	1.9	8.6	1.5	-4.69	.000	-
14	4.7	1.3	7.5	2.3	-4.58	.000	-
15	5.7	2.5	6.1	1.8	-0.60	.550	

Table 2: Correctness Score Comparison of Languages by Question

Hypothesis H3: SQL subjects took significantly longer to answer questions than did USL subjects (r=.303, p=.000, n=1042), providing support for H3.

Hypothesis H4: USL subjects with no training performed very poorly in the exam (see Table 3). Only 4.1% of their queries were "essentially correct" (44.6% for trained USL subjects). They tended to answer questions by describing algorithmic procedures, rather than directly querying the database; thus they consistently stayed outside the language rules (Welty-correctness code 'XF'). For example, an answer of a subject was:

> "Please get id of companies and individuals that have donated more than 20000 in 1981 from the donations table. Take the id and match up with the alumni or company from the personal information of appropriate tables. List last name, City, State and Zip of both alumni and companies."

CORRECTNESS	Mean	S. D.
Trained USL	6.89	2.31
Untrained USL	2.83	1.89
WELTY-SCALE	**Mean**	**S. D.**
Trained USL	5.60	2.39
Untrained USL	3.30	1.15

Table 3: Performance of USL subjects

While no significant differences were found between mean values of clarity, solution strategy, and perceived complexity and treatment, a significant association was found between these variables and test score (clarity-score: r=.238, p=.000, n=1044; solution strategy-score: r=.327, p=.000, n=1043; complexity-score: r=-.297, p=.000, n=1041). The negative association between perceived complexity and score suggests face validity because it would be expected that subjects would perform more poorly on the more complex questions.

Subjects who took a shorter amount of time answering a question tended to do better than subjects who took longer (score-time: r=-.142, p=.000, n=1040). It is likely that subjects who took a shorter time to answer a question were more certain about how to go about obtaining the answer.

Word Usage Results.

In total numbers, there were more unique words used in USL than in SQL for all queries. In contrast, there were more word occurrences used in SQL than in USL for all queries.

Tables 4 and 5 present the categories and number of words used in both languages (USL and SQL). For each language, words were categorized as nouns, verbs, etc. These categories were grouped in three major types: application-dependent words (TYPE I), application-independent words (TYPE II) and constant values (TYPE III). TYPE I words correspond to terms that must be defined for each new application (e.g. verbs, nouns, and adjectives). TYPE II words are predefined in the system core lexicon (e.g. prepositions, operators, articles, etc.). TYPE III words are the values that are stored in the database (e.g. numbers and proper names). Table 6 gives a summary of word usage.

Categories	Unique Words	Occurrences
TYPE I		
Verbs (non-imperative)	45	440
Nouns/Adjectives	101	1592
TYPE II		
Verbs (imperative)	8	195
Pronouns	11	247
Operators	7	86
Comparatives	4	221
Connectives (conjunctives)	6	216
Articles	4	120
Prepositions	12	748
Modifiers	10	131
TYPE III		
Constant Values (#'s)	21	304
Constant Values (strings)	30	178
TOTALS	259	4478

Table: 4 Word Usage for USL

Categories	Unique words	Occurrences
TYPE I		
Verbs (non-imperative)	11	64
Nouns/Adjectives	79	3658
TYPE II		
Verbs (imperative)	2	349
Operators	5	132
Comparatives	4	538
Connectives (conjunctives)	2	231
Prepositions	10	483
TYPE III		
Constant values (#'s)	28	395
Constant values (strings)	39	231
TOTALS	180	6081

Table: 5 Word Usage for SQL

	TYPE I	TYPE II	TYPE III
UNIQUE WORDS:			
USL	56%	24%	12%
SQL	50%	13%	37%
ALL OCCURRENCES:			
USL	44%	45%	11%
SQL	28%	61%	11%

Table 6: Summary of Word Usage

In order to assess commonality of word usage among USL subjects the method of Miller [17] was used. For this, non-imperative verbs, nouns, and adjectives (TYPE I) were examined. A list of the top 25 words in frequency of use by all subjects was created. This list contained 6% of the total unique words and amounted to 49% of all word occurrences. Lists were also created containing the 25 most used words for each subject, and the commonality of words was assessed by contrasting all lists. On the average, each word used by a subject was also used by 9.2 other subjects (55 percent of the most commonly used words were shared). Furthermore, the top 5 words were shared by an average of 15.8 persons (93 percent). These results show an even greater degree of commonality than those observed in [17]. Miller observed that 44 percent of the 25 most commonly used words were shared, and that 62 percent of the top 5 words were shared. The difference is attributed to the higher degree of focus for this experiment (database querying versus procedure writing in Miller's experiment).

The application-dependent words that were used very infrequently were also examined. Words that occurred less than three times accounted for 44 percent of the unique words, but only accounted for 6.2 percent of all word occurrences. This means that they could be dropped without serious loss of overall performance.

QUESTION	Number of Words		Grammaticality		Naturalness	
	Mean	S.D.	Mean	S.D.	Mean	S.D.
1	13.8	5.1	4.2	1.0	4.5	1.0
2	26.0	1.6	3.1	1.6	3.1	1.7
3	31.7	9.5	1.6	0.9	1.5	1.7
4	20.1	5.6	3.4	1.4	3.5	1.5
5	11.2	4.8	3.8	1.1	4.1	1.3
6	17.9	3.7	2.8	1.3	2.5	1.6
7	37.8	11.3	3.4	1.5	3.3	1.7
8	27.6	7.5	3.2	1.4	3.0	1.5
9	18.9	4.7	2.5	1.3	2.1	1.4
10	25.7	6.6	3.5	1.4	3.8	1.4
11	9.7	3.3	3.8	1.2	3.6	1.4
12	12.9	5.3	2.4	0.9	1.8	1.1
13	23.1	4.5	3.1	1.4	2.9	1.6
14	25.0	6.3	2.8	1.3	2.6	1.5
15	17.3	3.4	3.1	1.4	3.2	1.5
Totals	21.2	5.5	3.1	1.3	3.0	1.5

Table 7: Frequency of Word Usage, Grammaticality, Naturalness.

Careful investigation of the strategies used by USL subjects revealed small differences. There were also small differences among USL subjects on the number of words used per question as shown in Table 7. The table also shows the mean values and standard deviations for the measures of grammaticality and naturalness of the subjects' answers. Even after USL training, the subjects had a strong tendency to write non grammatical answers (mean value of 3.1 in a scale from 1

to 5), an indication that NL systems should be more flexible in accepting English requests. Still, the subjects used fairly awkward and verbose expressions in attempting to meet the artificial restrictions of USL (mean value of 3.0 in a scale of 1 to 5 for naturalness). As an example, they would use: "Where does the alumnus whose first is john and whose last name is eastburn live?", rather than the more natural, "where does john eastburn live?".

In summary, USL subjects did not use many words. There was a high degree of commonality in strategy and in application-dependent word usage, and low frequency words were mainly synonyms to other commonly used words.

DISCUSSION OF RESULTS AND CONCLUDING REMARKS

The laboratory study results supported all four of the tested hypotheses. The study also gave the opportunity to explore some fundamental research questions, and we believe the results offer some evidence for the feasibility of using natural language for database queries with a restricted vocabulary.

No difference in subject performance was found on the basis of language type. The correctness and Welty category scores were found to be highly correlated. The finding of a longer answer time for SQL subjects is consistent with the finding that SQL subjects had an average query length that was substantially larger than the USL average length. If one assumes writing the query consumes a major proportion of query answer (response time) then it is reasonable to

expect that SQL subjects will take longer to answer a question than USL subjects.

The need for training to use natural language query systems which are quite demanding in restrictions (e.g. USL), may be a major reason of why USL subjects did not perform better than SQL subjects.

The results of the laboratory experiment are also quite consistent with previous findings in other portions of the ALP project and with previous research. The finding that all subjects scored high on the test suggests that both languages can be learned with a combination of instruction and practice. Using the same training method and scoring method (mean percentage of essentially correct scores) as Welty, the SQL treatment subject test scores are similar to those found by Welty [34] and Reisner [21]. Welty's SQL subjects (two tests, n=35 and n=39) had an essentially correct answer percentage of 67.0 and 59.5 on twenty questions of varying degrees of difficulty. This compares with the average essentially correct SQL subject score of 53.3 on fifteen questions of varying difficulty. In an earlier study similar to Welty's, Reisner's SQL subjects had a percentage of essentially correct scores of 72 (n=64) using roughly the same scoring approach. Considering differences in subjects, training methods, material and time, and test content, the results of these studies are quite consistent.

We view the results of this laboratory study as a performance upper bound. That is, in real applications we would expect other factors, such as system loading, database size and complexity, operating system environment, the extent of networking, line

condition, and terminal type to reduce performance below what we and other researchers have observed in laboratory experiments. On the other hand, if a natural language query system provides constructive feedback to subjects, then learning may take place which could improve performance over that found in a laboratory setting.

In addition to testing hypotheses about the performance of the two languages (USL and SQL), this laboratory study allowed for the investigation of other fundamental research questions. These questions address the philosophy and structure of NL systems. The results here were positive. It seems possible to impose a fairly small vocabulary in such systems, since subjects did not use very many words and tended to use some common words very frequently. Also, after training, subjects used similar strategies in answering questions.

Acknowledgments. We take the opportunity to thank our colleague Margrethe Olson for many constructive comments in reviewing earlier versions of this paper. Also, several suggestions of the anonymous referees are gratefully acknowledged.

References

1. Artificial Intelligence Corp., "Intellect Query System", Reference Manual, (1982).

2. Bates, M., Bobrow, R.J., "A Transportable Natural Language Interface for Information Retrieval", Proceedings of the 6th Intl ACM SIGIR Conf, ACM special interest group on information retrieval and American Society for Information Science, Washington, D.C., June, 1983.

3. Codd, E.F., "Seven Steps to Rendezvous with the Casual User", Data Base Management, Klimbie and Koffeman (Eds), North-Holland, (1974), pp.179-199.

4. Damerau, F.J., "The Transformational Question Answering (TQA) System Operating Statistics", IBM Research Report, RC 7739, (1979).

5. Denny, G.H., "An Introduction to SQL, A Structured Query Language", Tech. Rep. RA93, IBM Research Lab, San Jose, (1977).

6. Egly, D.G., and Loebner, E.E., "Evidence for Natural Language Use to Decipher REL English Database Access", Hewlett-Packard Company, CSL-82-6, February (1982).

7. Grosz, B., et al, "TEAM: A Transportable Natural Language System", Technical Report no.263, SRI Artificial Intelligence Center, April, (1982).

8. Harris, L.R., "User Oriented Data Base Query with the ROBOT Natural Language Query System", Int. J. of Man-Machine Studies, 9, (1977).

9. Jarke, M., and Vassiliou, Y., "Choosing a Database Query Language", submitted for publication, November, (1982).

10. Kelly, M.J., and Chapanis, A., "Limited Vocabulary Natural Language Dialogue, International Journal of Man-Machine Studies, 9, (1977).

11. Krause, J., "Preliminary Results of a User Study with the 'User Specialty Languages' System, and Consequences for the Architecture of Natural Language Interfaces", IBM Heidelberg Scientific Center TR 79.04.003, (1979).

12. Lehmann, H. 'Interpretation of Natural Language in an Information System', IBM Journal of Research and Development, vol.22, no.5, September, (1978).

13. Lehmann, H., Ott, N., and M. Zoeppritz, "User Experiments with Natural Language for Data Base Access", Proceedings of 7th International Conference on Computational Linguistics, Bergen, 1978.

14. Malhotra, A., "Design criteria for a Knowledge-based English Language System for Management", MIT Project MAC, Cambridge MA, 1975.

15. Malhotra, A. and I. Wladawsky, "The Utility of Natural Language Systems," Research Report #RE5739, IBM T. J. Watson Research Center, Yorktown Heights, NY, (1975).

16. Miller, H.G., Hershman, R.L., and Kelly, R.T., "Performance of a Natural Language Query System in a Simulated Command Control Environment", Naval Electronics System Command, (1978).

17. Miller, L.A., "Natural Language Programming: Styles, Strategies, and Contrasts", "IBM systems Journal, 20, 2, (1981).

18. Ott, N. and M. Zoeppritz, "USL - An Experimental Information System Based on Natural Language," in L. Bolc ed. Natural Language Based Computer Systems, Macmillan, London., (1979).

19. Petrick, S.R., "On Natural Language Based Computer System", IBM Journal of Research and Development, 20, 4, July, 1976.

20. Reisner, P., "Human Factors Evaluation of Two Data Base Query Languages: SQUARE and SEQUEL", Proceedings of NCC, Vol. 44, (1975).

21. Reisner, P., "Use of Psychological Experimentation as an Aid to Development of a Query Language", IEEE Transactions of Software Engineering, SE-3, 3, (1977), pp.218-229.

22. Reisner, P., "Human Factors Studies of Database Query Languages: A Survey and Assessment", ACM Computing Surveys, 13, (1981), pp.13-32.

23. Schwartz, S.P., "Problems with Domain-Independent Natural Language Database Access Systems", Proceedings of the 20th Annual Meeting of the ACL, Toronto, Ont., Canada, June, (1982).

24. Shneiderman, B., "Improving the Human Factor Aspect of Database Interactions", ACM Transactions on Database Systems, 3, (1978).

25. Shneiderman, B., Software Psychology, Withrop, Cambridge/Mass., (1980).

26. Small,D.W., and Weldon,L.J., "The Efficiency of Retrieving Information From Computers Using Natural and Structured Query Languages", Rep. SAI-78-655-WA, Science Applications, September, (1977).

27. Stohr, E.A, J.A.Turner, Y.Vassiliou, N.H.White, "Research in Natural Language Retrieval Systems", 15th Ann.Hawaii Int.Conf. on System Sciences, Hawaii, (1982).

28. Tennant, H., "Experience with the Evaluation of Natural Language Question Answerers," Working Paper #18, Advanced Automation Group, Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, Urbana, Il., (1979).

29. Thomas, J.C. and Gould, J.D., "A Psychological Study of Query by Example", Proceedings of NCC, Vol. 44, (1975).

30. Thompson, F. B., and B. H. Thompson, "Practical Natural Language Processing: The REL System as a Prototype," in Advances in Computers 13, M. Rubinoff and M. C. Yovitz, ed. Academic Press, New York., 1975.

31. Turner,J., Jarke,M., Stohr,T, Vassiliou,Y., and White,N., "Using Restricted Natural Language for Data Retrieval - A Field Evaluation", in Human Factors and Interactive Computer Systems, Y. Vassiliou (ed), ABLEX, Norwood, NJ, 1983.

32. Vassiliou, Y., and M. Jarke, "Query Languages - A Taxonomy", in Human Factors and Interactive Computer Systems, Y. Vassiliou (ed), ABLEX, Norwood, NJ, 1983.

33. Waltz, D.L., "An English Language Question Answering System for a Large Relational Database", Communications of the ACM, 21, (1978).

34. Welty, C., Stemple, D.W., "Human Factors Comparison of a Procedural and a Non-Procedural Query Language", ACM Transactions on Database Systems, (1981).

35. Woods, W.A., "Lunar Rocks in Natural English: Explorations in Natural Language Question Answering", Linguistic Structures Processing, Zampolli (ed.), North-Holland, (1977).