A CONTRAST BETWEEN COMPUTER AND HUMAN LANGUAGE UNDERSTANDING

CSE Technical Report 287

Eva L. Baker
Elaine L. Lindheim

Center for Technology Assessment
UCLA Center for the Study of Evaluation

May, 1988

This paper was presented at the 1988 Annual Meeting of the American Educational Research Association in New Orleans as part of a symposium titled "Understanding Natural Language Understanding."

The research reported herein was conducted with partial support from the Office of Naval Research, Defense Advanced Research Projects Agency, pursuant to grant number N00014-86-K-0395. However, the opinions expressed do not necessarily reflect the position or policy of this agency and no official endorsement by this agency should be inferred.

Please address inquiries to: CSE Dissemination Office, UCLA Graduate School of Education, 405 Hilgard Avenue, Los Angeles, California, 90024-1521

Goals and Objectives

This report will present progress in exploring an approach to the evaluation of intelligent computer systems, in particular in the area of natural language (NL) understanding. Our overall project strategy is to develop a multidimensional system to evaluate both qualitative and quantitative elements of natural language computer programs. The reasons for this project are threefold. First, it is difficult for program managers and potential users of systems to get clear and consistent indicators of system performance and improvement in other than very technical terms. The evaluation of such systems, while a hot topic, has proceeded unsystematically and in general without regard to the long history in evaluation and measurement shared by the social sciences. Last, as a research enterprise, we are interested in understanding how and how much of computer programs purported to model intelligence can be referenced back to the performance of real people. We intend to try to apply techniques from psychometrics to evalute natural language programs.

The focus of this document is our research in relating human performance measures to NL implementations. Although our work extends to other areas, i.e., vision and expert systems, we believe we have the best chance of success in the NL area for two reasons. First, the natural language area is one of the most well developed in the AI community with literally scores of programs aping language understanding. Second, from the educational measurement side, there is a long history and extensive set of testing approaches related to reading comprehension. Our task is essentially to determine if there are unions between these two traditions that will help us describe and assess natural language implementations in terms of measured human language proficiencies.

Technical Approach

In the simplest terms, we intend to norm a given NL system's performance on a sample of people. We have begun to create tests that measure the language functions of target programs and to benchmark systems in terms of the characteristics and abilities of the human performance. It is our intention to apply our approach to a sample of natural language implementations. A typical NL implementation used for research might consist of a discrete piece of text, perhaps a description of a common scene (Dyer, 1983). The goal of the NL developer is to demonstrate that the computer can understand both literally and inferentially what has happened. The mode of demonstration is asking questions of the system. In order to respond, complex rules are programmed describing explicitly the context needed to answer the questions. In this simplest of cases, our benchmarking approach would require the following:

1. Develop domain specifications appropriate to generate questions about the text segment.

2. Generate test items appropriate to the text.

3. Create a measure consisting of the NL developer's questions and our own and administer to "norming" or referent groups.

4. Describe NL system performance in terms of the group whose responses are most comparable to those of the system.

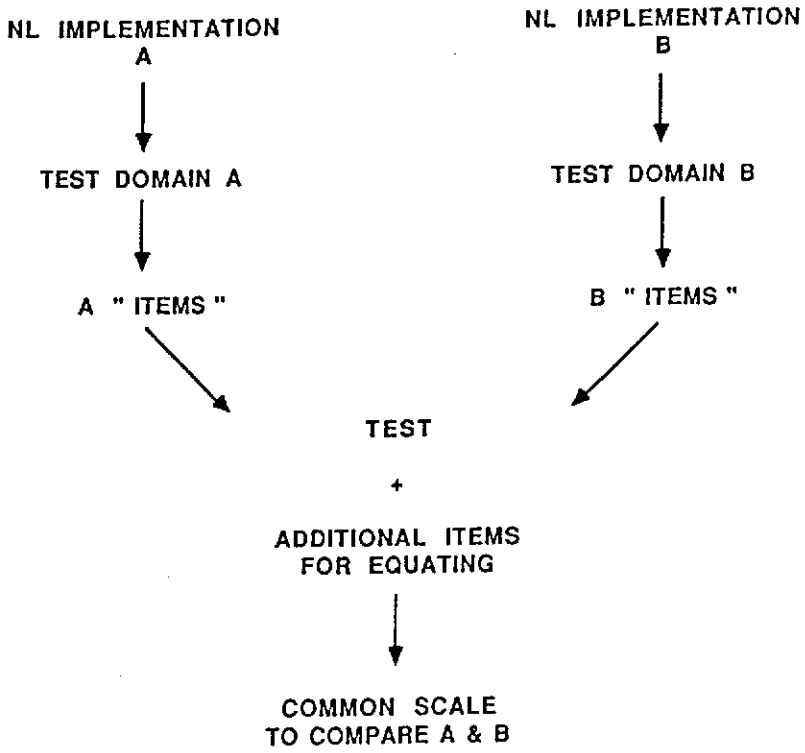
We are also testing the feasibility that comparisons among systems can be made. After completing the above task for each of two separate NL programs, if comparisons were desired, an additional set of steps would follow (see Figure 1):

1. One or more constructs would be posited.

2. Anchor items would be developed and administered to the same norming groups.

3. Analyses to assess the equating options would be conducted.

Figure 1
Benchmark Comparison Example



Clearly, this approach appears to gloss over some important differences between systems and people. For example, we have not decided (nor is it really feasible) to measure explicitly important other language performances the comparison group can accomplish in addition to those targeted by the system. People are obviously infinitely more creative and proficient in language than any system yet or to be devised. Yet, a quick reading of our project might imply that we will infer that system peformance equals human performance. We will not. Conversely, there will be aspects of system performance clearly superior to what people can do -- perfect reliability for one example not familiar to psychometricians. Our approach is exploratory and its utility will depend upon how sensible and understandable our comparisons will be.

The Natural Language System: IRUS

For the remainder of the report, we will provide specific details of our progress in developing human performance benchmarks for a specific natural language system.

Unlike the simple text based example above, the actual system to which we applied our methods provides us with a greater challenge. The program under study is a natural language query system. Essentially such a program permits the user to ask questions in regular English prose to another computer program, perhaps a database or an expert system. The natural language system, IRUS (Bates, Stallard, & Moser, 1985.), is an interface between the user and the set of information desired for access, and provides a rapid, natural and convenient method for obtaining information. The particular interface we are assessing has been designed, at least so far, to serve as a general purpose interface to a broad range of databases and expert systems. It is a basic syntactic shell that needs to be filled with specifics in order to work. To use IRUS, it must be specifically adapted to a designated database. The particular semantics (content) of the database or expert system must be translated into rules used by IRUS. Here are a sample of queries that IRUS can deal with, demonstrated in two different domains.

Table 1
IRUS in Two Domains

IRUS in a library science domain:

"Of the books on Artificial Intelligence, how many have been classified as textbooks?"

"Have there been as many requests for books about medicine this year as we planned for in our budget?"

"Which organizations that we receive reports from have responded to either of our recent questionaires?"

IRUS in the domain of Navy ships:

"List the number of ships that are deployed in the Indian Ocean."

"What's the name of the commander of Frederick?"

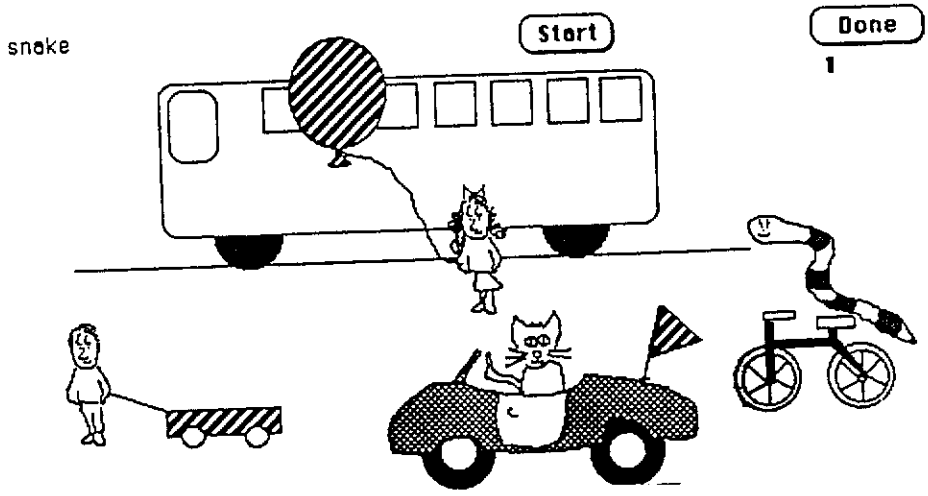
"What is Vinson's current course?"

Our knowledge of what IRUS can do has come from a system test of IRUS, where the system successfully answered a series of 165 questions. We have taken these questions, classified them into semantic and syntactic categories, and developed a set of test specifications designed to measure human ability to understand these questions.

Because IRUS is always embedded in a specific domain, e.g., ships status, it is clear that our measurement approach needed to separate out the understanding of the question from the ability to provide the answer. Clearly, making comparisons based on the correctness of answers about the location of Navy ships makes no sense. Because we believed that many of the IRUS query types could be answered by very young children, we decided to develop a measure that would provide for children a very simple database - one consisting of animals, people, houses, their attributes, and positions.

The measure includes a pretest that determines whether students understand the elements in the database. The pretest is shown in Figure 2. By screening out examinees who cannot identify the database elements, we are able to infer that students' selection of the correct answer is based upon their understanding of the question. In our study, unlike the real IRUS applications, the databases function only to permit us to assess language function.

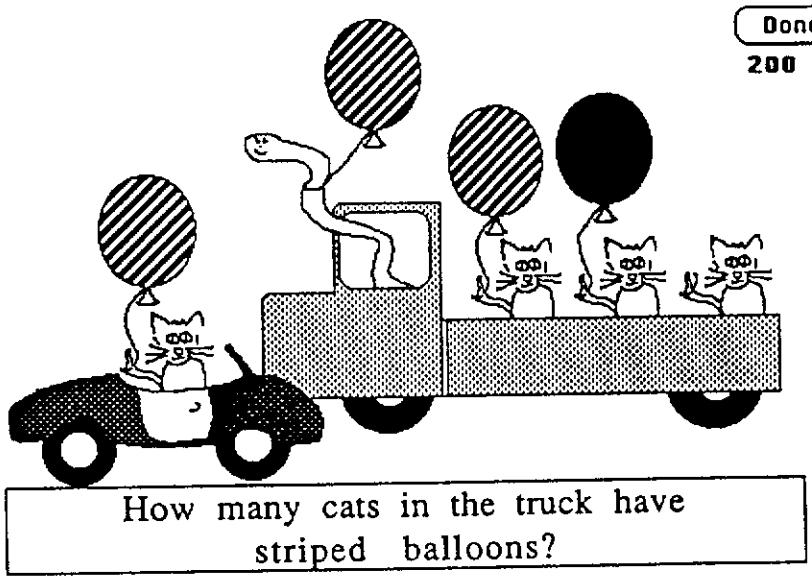
Figure 2
IRUS Pretest



IRUS Vocabulary Pretest. Students are shown a prompt (e.g., "Point to the snake") and asked to respond.

We have included copies of some of the test items presented to students for our prototype test (see Figures 3 through 6). The test was implemented in Hypercard and administered on Macintosh SE computers.

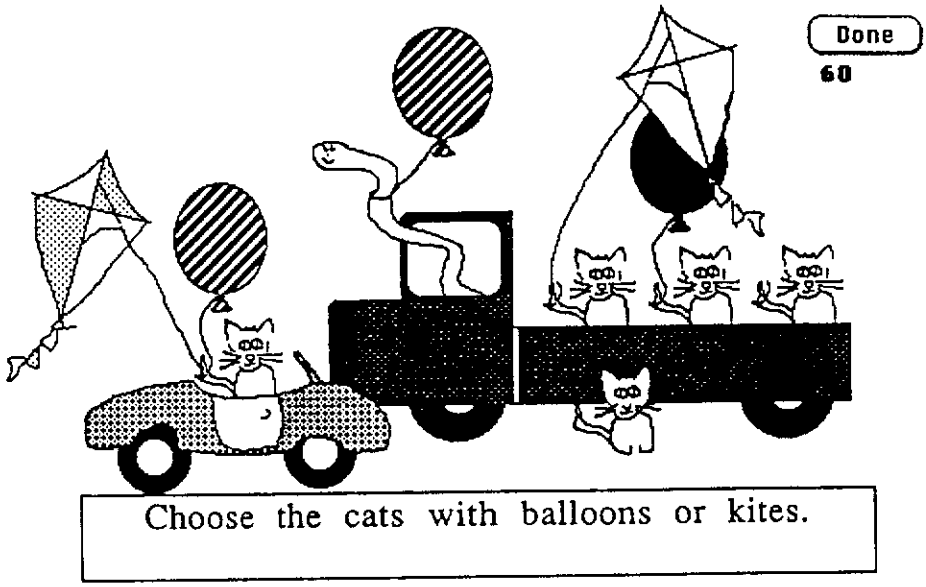
Figure 3
Test Item



Comparable IRUS queries:

How many ships in the Third Fleet are C-3?
How many of the ships in Indian Ocian are C-5?

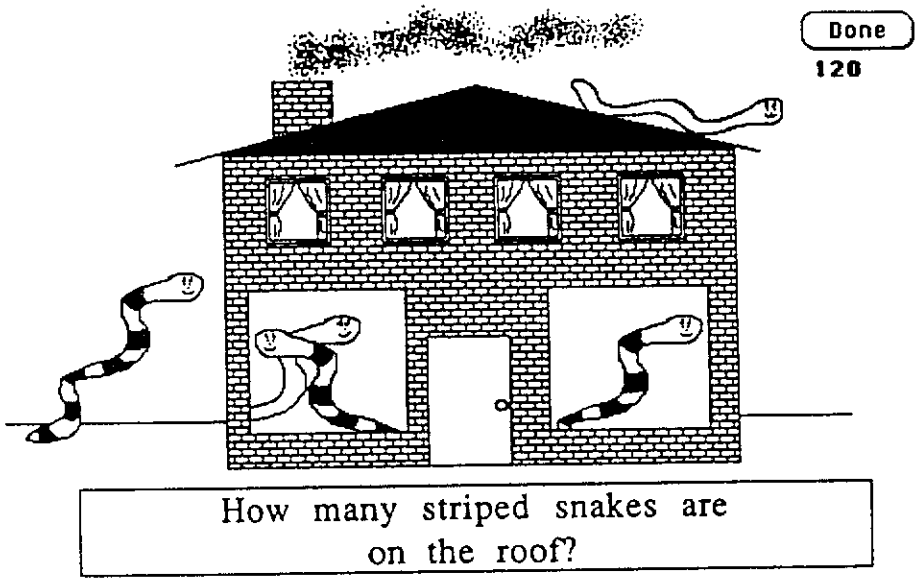
Figure 4
Test Item



Comparable IRUS query:

List the ships that are C4 or that are C5.

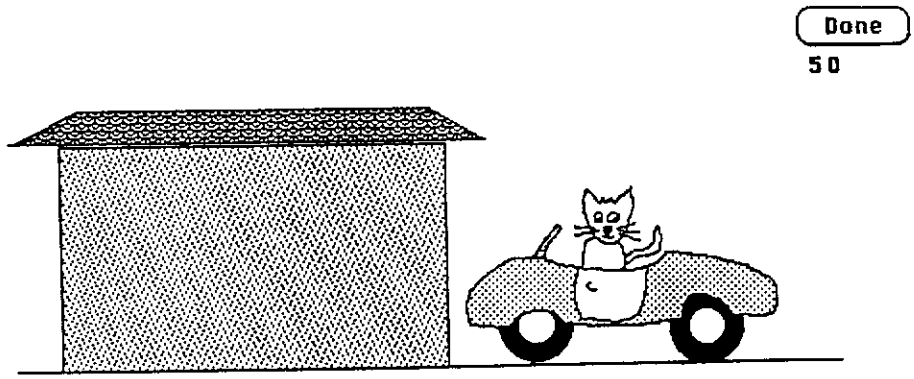
Figure 5
Test Item



Comparable IRUS queries:

How many harpoon capable ships are in PACFLT?
How many US ships are in the Indian Ocean?

Figure 6
IRUS Pretest



Comparable IRUS queries:

Is the Kennedy in port?
Is Vinson in San Diego?

Procedures

In order to determine the appropriate language understanding level at which to administer the IRUS test, we have piloted the test with early elementary school and preschool students. Those students who are reading at a second grade level or higher read the questions themselves and the test administrator reads aloud the queries for younger students.

Depending upon the type of query, examinees answer with either an oral response or by pointing to the answer on the computer screen. When the examinee response orally, the administrator types the answer on the screen so that it is entered in the computer transcript of the test. For example, a student might answer the query, "How many cars have striped flags", by saying "four." The administrator types "four" into the the transcript. When the examinee points to an answer on the screen, the administrator uses the mouse to highlight the student's choice. (For example, a student might answer the query "Choose the cats with striped balloons" by pointing to any cats on the screen that fulfilled the requirement. The administrator would click on each animal identified by the student.)

In a talk-aloud procedure designed to validate students' understanding of the questions they have answered, students are asked to explain their responses to the more complex queries. After their response has been entered in the computer, the administrator asks, "Why did you say '. . .'" or "How did you know that '. . .' was the answer?" These responses are tape recorded for analysis in conjunction with the test transcript.

Results

Results to date indicate the following:

1. Students reading at or above a second grade level generally can recognize all of the elements in the database when those elements are presented in the pretest.

2. Students reading at or above a second grade level have no difficulty in reading aloud the test queries. Some students skip over words and even paraphrase when reading aloud (for example, one student substituted the word "bike" for "bicycle"), but these alterations do not stand in the way of the examinees providing the correct answer.

3. Some students reading at or above a second grade level prefer to process the query silently, rather than reading it aloud, even when reminded that the test administration procedures call for them to read aloud.

4. Students reading at or above a second grade level have difficulty with those queries that are more than 9 words long and that contain more than one delimiter, when the relationship between delimiters is expressed by either the conjunction "and" or the conjunction "or."

5. Students reading at or above a second grade level often answer a query with a response that is literally incorrect but pragmatically valid. For example, when asked, "How many cars have striped flags?", they may respond by pointing to all the cars that meet the stipulation rather than answering with a specific number.

6. Students reading at or above a second grade level provide very little additional information when asked to explain their answers to test queries. Typical answers to thequestion, "How did you know that was the correct answer?", are to point to specific details in the picture or to state "because that's the answer."

Next Steps

To complete the IRUS test program, we must identify by semantic class reliable levels of performance for identifiable student groups. This means that we must test samples of students at higher and lower levels of performance than our first trials.

Since our task involves providing understandable benchmarks, we must be sure that we are accurately describing our sample of students. Our first set of student descriptors is very gross: grade level and age. It may be necessary to develop and refine student descriptors to determine which kinds of people have difficulty with the queries and why. For example, performance on standard language proficiency measures may be used as a more fine grained description of our student samples.

We are in the process of refining our test formats and plan to use existing and planned think-aloud protocols to help us develop alternative approaches for assessing the competencies of IRUS. Our approach should help us in any subsequent evaluation of NL interfaces. We are now in the process of identifying other NL programs for testing and hope to complete at least two more tests in calendar 1988. At that point, we will be able to undertake equating studies that will allow us to compare disparate NL programs.

Last, we also must carefully assess the utility of this approach to determine whether it provides members of the research and development communities a useful way to characterize natural language programs.

References

Bates, M., Stallard, D., & Moser, M. (1985). The IRUS transportable natural language database interface. In *Expert Database Systems*. Menlo Park, CA: Cummings Publishing Company.

Dyer, M.C. (1983). *In-depth understanding*. Cambridge, MA: MIT Press.