Letters to the Editor

Re Ballard on the Need for Careful Description[1]

I think, like the reviewer of Bruce Ballard's previous paper, that he wants the moon. However, as one who has tried, for the purposes of lecturing, to extract concrete system descriptions with worked-through examples from published material, I think he is right to call for better standards. Alan Bundy (1981) has made a similar appeal for AI in general. It's important, in particular, to appreciate that raising the standard of reporting raises the standard not only of the reader's work but that of the writer's: anyone who is obliged to provide a coherent account of a set of experiments soon discovers which ones he hasn't done and needs to do, sometimes forthwith before continuing writing.

The problem Ballard does not face, and should say something about, is the scale impact of his proposal: providing everything called for in useful, if not exhaustive, detail, is liable to generate very long papers. One can indeed plough through whole theses intended in principle, and not conspicuously failing in practice, to provide what Ballard calls for, and still not find sufficient evidence of what has been done and, more importantly, how it has been done. How much grammar, and more significantly, how much dictionary, should you put in to support your description and claims for performance?

Ballard would give much more meat to his case if he provided some concrete examples of papers he feels comes closest to what he is asking for, with some comments on their successes and failures. How well, to take some random examples, do Waltz's (1978) PLANES paper, Erman et al. (1980) on Hearsay-II, or Warren and Pereira's (1982) Chat-80 measure up, or, on a larger scale, Woods's (1972) LUNAR report?

But perhaps the correct response to Ballard's suggestion is to ask him to take some system and provide the kind of account of it he is looking for. Show us the way, friend.

Karen Sparck Jones
Computer Laboratory
University of Cambridge
Corn Exchange Street
Cambridge CB2 3OG ENGLAND

[1] Letter to the Editor, *AJCL* 9(1): 23-24.

References

Bundy, A. 1981 *AISB Quarterly* 40(1): 226-228.
Erman, L.D. et al. 1980 *ACM Computing Surveys* 12: 213-253.
Waltz, D.L. 1978 *Communications of the ACM* 21: 526-539.
Warren, D.H.D. and Pereira, F.C.N. 1982 *American Journal of Computational Linguistics* 8: 110-122.
Woods, W.A. et al. 1972. Report 2378, Bolt Beranek and Newman Inc.

On the Need for Studying Ill-Formed Input

The experiment described by Fineman (1983) provides important and useful information about the extent and nature of ill-formed input to natural language systems. A very low level of ill-formed input was found in this experiment. This result contrasts with the much higher level of ill-formed input found in the experiment described by Eastman and McLean (1981). A consideration of the different experimental situations reveals many factors that could account for this difference. (The experiment described by Fineman will be referred to as the Duke experiment; that described by Eastman and McLean, as the Florida State University (FSU) experiment.)

Many more restrictions were placed upon the input requested from the Duke subjects than from the FSU subjects. Also, the Duke subjects were provided with more opportunities to learn about the capabilities of the system. Both of these factors would be expected to result in a lower rate of ill-formed input.

Experimental Goals. The goal of the Duke experiment was to evaluate and guide the design of a proposed natural language system. The goal of the FSU experiment was to compare requests posed to a simple data base by users with different levels of experience with computers and with the example data base.

System Interaction. The Duke experiment used simulated voice-driven input; subjects were asked to use discrete speech or slow connected speech. Less constrained speech might have contained more errors. The FSU experiment used sentences handwritten on a questionnaire. This difference in input method would be expected to lead to different results. Also, some of the errors found in the FSU experiment, such as misplaced apostrophes and misspellings, would not be relevant in a voice input system.

Feedback. Simulated system response was provided to users in the Duke experiment. Thus they had an opportunity to learn about the system and to modify their behavior. Mistakes would be less like to be repeated.

No feedback was provided in the FSU experiment. Some subjects made the same type of mistake in all of their queries. Had they had the opportunity to receive feedback, they might not have repeated those particular errors. At a later time, some of the subjects used an implementation of the system, and several commented that they found it relatively easy to learn what the system could handle and to adapt to it.

Number of Subjects. The Duke experiment used 15 subjects; the FSU experiment used 231.

Subject Population. The subjects used in the FSU experiment were FSU students. The population from which the Duke subjects were drawn was not identified. It is quite possible that differences in ability and training between the groups were contributing factors.

Session Length. Subjects interacted with the system for an hour in the Duke experiment. The FSU subjects spent about 15 minutes reading and responding to a questionnaire.

Number of Sentences Collected. About 1,600 sentences were collected in the Duke experiment; 693 (3 per subject) were collected in the FSU experiment. Far more information was collected from each subject in the Duke experiment.

Vocabulary Restriction. The vocabulary in the Duke experiment was restricted to about 50 words. No restriction was placed upon vocabularly in the FSU experiment, and the sample queries collected contained a few hundred words. A vocabulary of 50 words is unlikely to be able to handle a subject domain of reasonable size. More errors might have been made with a larger vocabulary.

Sentence Restriction. In the Duke experiment, each sentence had to begin with an imperative verb. No restriction was placed on the sentences in the FSU experiment, but the subjects were told that the system could only handle "simple" sentences. One of the four examples they were given began with an imperative verb; the others were questions.

Of the 693 queries collected in the FSU experiment, approximately 15% (102) began with an imperative verb. No obvious differences in the kinds of ill-formed input found in the different types of sentences were noted. All of the requests were for retrieval of information; there were no requests for other system actions, such as printing, for which imperative commands might be more frequently used.

Domain. The topic in the Duke experiment was the "office domain"; that in the FSU experiment was a simple personnel data base.

Training. Subjects in the Duke experiment were provided with a short tutorial. Subjects in the FSU experiment were provided with a few examples.

The results reported in these two experiments are complementary rather than conflicting. They illustrate the danger of generalizing from one type of natural language system to another and the need for studies under operational conditions. It is a mistake to include capabilities for handling ill-formed input that will not be needed; it is also a mistake to omit such capabilities when they will be needed.

Not only do we need to know how to handle different types of ill-formed input, we also need to know the conditions under which such features need to be included in systems. It is not to be expected that the latter type of information can be obtained from any single experiment, no matter how well designed or conducted. Patterns are likely to emerge only from a variety of experiments conducted under different conditions.

C. M. Eastman
Department of Computer Science
and Engineering
Southern Methodist University
Dallas, TX 75275

D. S. McLean
IBM Corporation
P.O. Box 1328
Boca Raton, FL 33432

References

Eastman, C.M. and McLean, D.S. 1981 On the Need for Parsin Ill-Formed Input. *American Journal of Computation Linguistics* 7(4): 257.

Fineman, L. 1983 Questioning the Need for Parsing Ill-Formed Input. *American Journal of Computational Linguistics* 9(1): 22.