PRELIMINARY EVALUATION OF THE VOYAGER SPOKEN LANGUAGE SYSTEM*

Victor Zue, James Glass, David Goodine, Hong Leung,
Michael Phillips, Joseph Polifroni, and Stephanie Seneff

Spoken Language Systems Group
Laboratory for Computer Science
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

ABSTRACT

VOYAGER is a speech understanding system currently under development at MIT. It provides information and navigational assistance for a geographical area within the city of Cambridge, Massachusetts. Recently, we have completed the initial implementation of the system. This paper describes the preliminary evaluation of VOYAGER, using a spontaneous speech database that was also recently collected.

INTRODUCTION

One of the important factors that have contributed to the steady progress of the Strategic Computing Speech program has been the establishment of standardized performance evaluation procedures [1]. With the use of common databases and metrics, we have been able to objectively assess the relative merits of different approaches and systems. These practices have also had a positive influence on the natural language community in that databases and rigorous evaluation procedures for natural language systems are beginning to emerge. As we move towards combining speech recognition and natural language technology to achieve speech understanding, it is essential that the issue of performance evaluation again be addressed early on, so that progress can be monitored and documented. Since the Spoken Language Systems program is in its infancy, we do not as yet have a clear idea of how spoken langauge systems should be evaluated. Naturally, we should be able to benefit from hands-on experience with applying some candidate performance measures to working systems. The purpose of this paper is to document our experience with the preliminary evaluation of the VOYAGER system currently under development at MIT, so that we may contribute to the evolutionary process of defining the appropriate evaluation measures.

VOYAGER is a speech understanding system that can provide information and navigational assistance for a geographical area within the city of Cambridge, Massachusetts. The components of the system are described in a companion paper [2]. To evaluate VOYAGER we made use of a spontaneous speech database that we have recently collected consisting of nearly 10,000 sentences from 100 speakers. The database is described in another companion paper [3].

EVALUATION ISSUES

We believe that spoken language systems should be evaluated along several dimensions. First, the *accuracy* of the system and its various modules should be documented. Thus, for example, one can measure a given system's phonetic, word, and sentence accuracy, as well as linguistic and task completion accuracy. Second, one must measure the *coverage* and *habitability* of the system. This can be applied to the lexicon, the language model, and the application back-end. Third, the system's *flexibility* must be established. For

*This research was supported by DARPA under Contract N00014-89-J-1332, monitored through the Office of Naval Research.

example, how easy is it to add new knowledge to the system? How difficult is it to port the system to a different application? Finally, the *efficiency* of the system should be evaluated. One such measure may be the task completion time.

Whether we want to evaluate the accuracy of a spoken language system in part or as a whole, we must first establish what the *reference* should be. For example, determining word accuracy for speech recognizers requires that the reference string of words first be transcribed. Similarly, assessing the appropriateness of a syntactic parse presupposes that we know what the correct parse is. In some cases, establishing the reference is relatively straightforward and can be done almost objectively. In other cases, such as specifying the correct system response, the process can be highly subjective. For example, should the correct answer to the query, " Do you know of any Chinese restaurants?" be simply, "Yes," or a list of the restaurants that the system knows?

It is important to point out that at no time is a human totally out of the evaluation loop. Even for something as innocent as word accuracy, we rely on the judgement of the transcriber for ambiguous events such as "where is," versus "where's," or "I am" versus "I'm." Therefore, the issue is *not* whether the reference is obtained objectively, but the degree to which the reference is tainted by subjectivity.

The outputs of the system modules naturally become more general at the higher levels of the system since these outputs represent more abstract information. Unfortunately, this makes an automatic comparison with a reference output more difficult, both because the *correct* response may become more ambiguous and because the output representation must become more flexible. The added flexibility that is necessary to express more general concepts also allows a given concept to be expressed in many ways, making the comparison with a reference more difficult.

To evaluate these higher levels of the system, we will either have to restrict the representation and answers to be ones that are unambiguous enough to evaluate automatically, or adopt less objective evaluation criteria. We feel it is important not to restrict the representations and capabilities of the system on account of an inflexible evaluation process. Therefore, we have begun to explore the use of subjective evaluations of the system where we feel they are appropriate. For these evaluations, rather than automatically comparing the system response to a reference output, we present the input and output to human subjects and give them a set of categories for evaluating the response. At some levels of the system (for example evaluating the appropriateness of the response of the overall system) we have used subjects who were not previously familiar with the system, since we are interested in a user's evaluation of the system. For other components of the system, such as the translation from parse to action, we are interested in whether they performed as expected by their developers, so we have evaluated the output of these parts using people familiar with their function.

In the following section, we present the results of applying various evaluation procedures to the VOYAGER system. We don't profess to know the answers regarding how performance evaluation should be achieved. By simply plunging in, we hope to learn something from this exercise.

PERFORMANCE EVALUATION

Our evaluation of the VOYAGER system is divided into four parts. The SUMMIT speech recognition system is independently evaluated for its word and sentence accuracy. The TINA natural language system is evaluated in terms of its coverage and perplexity. The accuracy of the commands generated by the back end is determined. Finally, the appropriateness of the overall system response is assessed by a panel of naive subjects. Unless otherwise specified, all evaluations were done on the designated test set [3], consisting of 485 and 501 spontaneous and read sentences, respectively, spoken by 5 male and 5 female subjects. The average number of words per sentence is 7.7 and 7.6 for the spontaneous and read speech test sets, respectively.

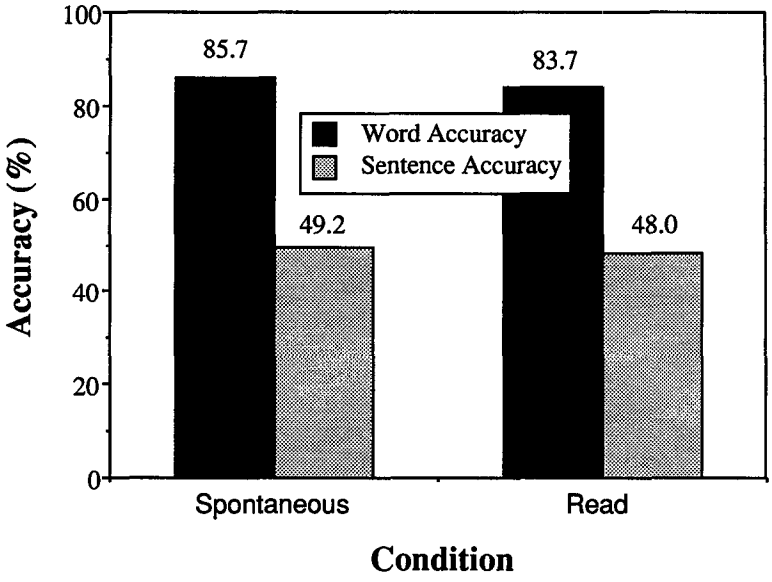

Figure 1: Word and sentence accuracy for the spontaneous and read speech test sets.

SPEECH RECOGNITION PERFORMANCE

The SUMMIT speech recognition system that we evaluated is essentially the same as the one we described during the last workshop [4], with the exception of a new training procedure as described elsewhere [2]. Since the speech recognition and natural language components are not as yet fully integrated, we currently use a word-pair grammar to constrain the search space. The vocabulary size is 570 words, and the test set perplexity and coverage are 22 and 65% respectively.[1] Figure 1 displays the word and sentence accuracy for SUMMIT on both the spontaneous and read speech test sets. For word accuracy, substitutions, insertions and deletions are all included. For sentence accuracy, we count as correct sentences where all the words were recognized correctly. We have included only those sentences that pass the word-pair grammar, following the practice of past Resource Management evaluations. However, overall system results are reported on *all* the sentences. For spontaneous speech, we broke down the results into three categories: sentences that contain partial words, sentences that contain filled pauses, and uncontaminated sentences. These results are shown in Figure 2. Since we do not explicitly model these spontaneous speech events, we expected the performance of the system to degrade. However, we were somewhat surprised at the fact that the read speech results were very similar to the spontaneous speech ones (Figure 1). One possible reason is that the speaking rate for the read speech test set is very high, about 295 words/min compared to 180 words/min for the spontaneous speech and 210 words/min for the Resource Management February-89 test set. The read speech sentences were collected during the last five minutes of the recording session. Apparently, the subjects were anxious to complete the task, and we did not explicitly ask them to slow down.

NATURAL LANGUAGE PERFORMANCE

Following data collection, TINA's arc probabilities were trained using the 3,312 sentences from the designated training set [5]. The resulting coverage and perplexity for the designated development set are shown

[1]The vocabulary in this case is larger than that for the entire system. The latter is the intersection of the recognition component's vocabulary with that of the natural language component.

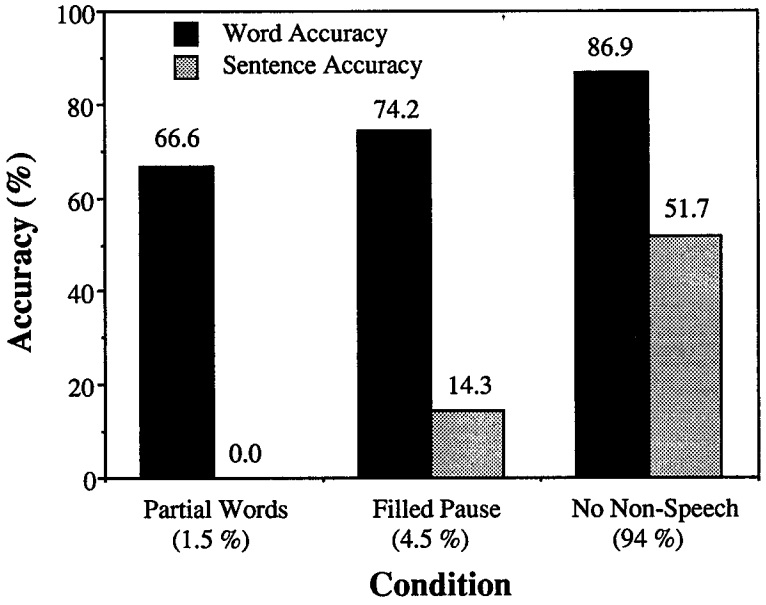

Figure 2: Breakdown of word and sentence accuracy for the spontaneous speech test sets, depending on whether the sentences contain false starts or filled pauses.

in the top row of Table 1. The left column gives the perplexity when all words that could follow a given word are considered equally likely. The middle column takes into account the probabilities on arcs as established from the training sentences. The right column gives overall coverage in terms of percentage of sentences that parsed.

Examination of the training sentences led to some expansions of the grammar and the vocabulary to include some of the more commonly occurring patterns/words that had originally been left out due to oversight. These additions led to an improvement in coverage from 69% to 76%, as shown in Table 1, but with a corresponding increase in perplexity. This table also shows the performance of the expanded system on the *training* set. The fact that there is little difference between this result and the result on the development set suggests that the training process is capturing appropriate generalities. The final row gives perplexity and coverage for the test set. The coverage for this set was somewhat lower, but the perplexities were comparable.

Note also that perplexity as computed here is an upper bound measurement on the actual constraint provided. In a parser many long-distance constraints are not detected until long after the word has been incorporated into the perplexity count. For instance, the sentence "What does the nearest restaurant serve?" would license the existence of "does" as a competitor for "is" following the word "what." However, if "does" is actually substituted for "is" incorrectly in the sentence "What is the nearest restaurant?" the parse would fail at the end due to the absence of a predicate. It is difficult to devise a scheme that could accurately measure the gain realized in a parser due to long-distance memory that is not present in a word-pair grammar.

The above results were all obtained directly from the log file, as typed in by the experimenter. We also have available the orthographic transcriptions for the utterances, which included false starts explicitly. We ran a separate experiment on the test set in which we used the orthographic transcription, after stripping away all partial words and non-words. We found a 2.5% reduction in coverage in this case, presumably due to back ups after false starts.

Of course, we have not yet taken advantage of the constraint provided by TINA, except in an accept/reject mode for recognizer output. We expect TINA's low perplexity to become an important factor for search space reduction and performance improvement once the system is fully integrated.

Initial System			
	No-Prob	Prob	Coverage
Development Set:	20.6	7.1	69%
Expanded System			
	No-Prob	Prob	Coverage
Development Set:	27.1	8.3	76%
Training Set:	25.8	8.1	78%
Test Set:	26.0	8.2	72.5%

Table 1: Perplexity and coverage for TINA for a number of different conditions.

SYSTEM PERFORMANCE

VOYAGER's overall performance was evaluated in several ways. In some cases, we used automatic means to measure performance. In others, we used the expert opinion of system developers to judge the correctness of intermediate representations. Finally, we used a panel of naive users to judge the appropriateness of the responses of the system as well as the queries made by the subjects.

Automated Evaluation

VOYAGER's responses to sentences can be divided into three categories. For some sentences, no parse is produced, either due to recognizer errors, unknown words, or unseen linguistic structures. For others, no action is generated due to inadequacies of the back end. Some action is generated for the remainder of the sentences. Figure 3 show the results on the spontaneous speech test set. The system failed to generate a parse for one reason or another on two-thirds of the sentences. Of those, 26% were found to contain unknown words. VOYAGER almost never failed to provide a response once a parse had been generated. This is a direct result of our conscious decision to constrain TINA according to the capabilities of the back end.

For diagnostic purposes, we also examined VOYAGER's responses when orthography, rather than speech, was presented to the system, after partial words and non-words had been removed. The results are also shown in Figure 3. Comparing the two sets of numbers, we can conclude that 30% of the sentences would have failed to parse even if recognized correctly, and an additional 36% of the sentences failed to generate an action due to recognition errors or the system's inability to deal with spontaneous speech phenomena.

Even if a response was generated, it may not have been the correct response. It is difficult to know how to diagnose the quality of the responses, but we felt it was possible to break up the analysis into two parts, one measuring the performance of the portion of the system that translates the sentence into functions and arguments and the other assessing the capabilities of the back end. For the first part, we had two experts who were well informed on the functionalities in the back end assess whether the function calls generated by the interface were complete and appropriate. The experts worked as a committee and examined all the sentences in the test set for which an action had been generated. They agreed that 97% of the functions generated were correct. Most of the failures were actually due to inadequacies in the back end. For example, the back end had no mechanism for handling the quantifier "other" as in "any *other* restaurants," and therefore this word was ignored by the function generator, resulting in an incomplete command specification.

Human Evaluation

For the other half of the back end evaluation, we decided to solicit judgments from naive subjects who had had no previous experience with VOYAGER. We decided to have the subjects categorize both system

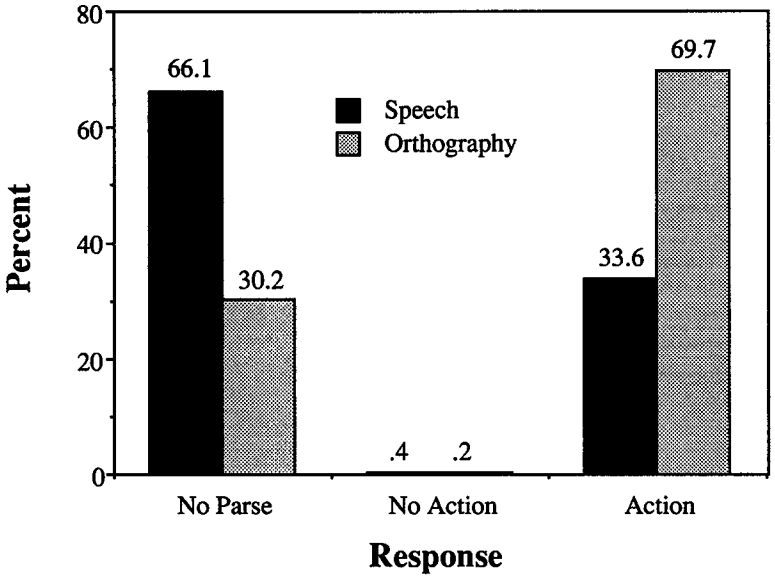

Figure 3: A breakdown of system performance for speech and orthographic input.

responses and user queries as to their appropriateness. System responses came in two forms, a direct response to the question if the system thought it understood, or an admission of failure and an attempt to explain what went wrong. Subjects were asked to judge answers as either "appropriate," "verbose," or "incorrect," and to judge error messages as either "appropriate" or "ambiguous." In addition, they were asked to judge queries as "reasonable," "ambiguous," "ill-formed," or "out-of-domain." Statistics were collected separately for the two conditions, "speech input" and "orthographic input." In both cases, we threw out sentences that had out-of-vocabulary words or no parse. We had three subjects judge each sentence, in order to assess inter-subject agreement.

Table 2 shows a breakdown (in percentage) of the results, averaged across three subjects. The columns represent the judgement categories for the system's responses, whereas the rows represent judgement categories for the user queries. A comparison of the last row of the two conditions reveals that the results are quite consistent, presumably because the majority of the incorrectly recognized sentences are rejected by the parser. About 80% of the sentences were judged to have an appropriate response, with an additional 5% being verbose but otherwise correct. Only about 4% of the sentences produced error messages, for which the system was judged to give an appropriate response about two thirds of the time. The response was judged incorrect about 10% of the time. The table also shows that the subjects judged about 87% of the user queries to be reasonable.

In order to assess the reliability of the results, we examined the agreement in the judgements provided by the subjects. For this limited experiment, at least two out of three subjects agreed in their judgements about 95% of the time.

SUMMARY

In this paper we presented some results on the preliminary evaluation of the VOYAGER system. As we have stated at the onset, we are entering into a new era of research, and we do not have a clear idea of how spoken language systems should best be evaluated. However, we have chosen to explore this issue along several dimensions. We have reached the conclusion that a totally objective measure of performance may not

	answer appropriate	answer verbose	error appropriate	error ambiguous	response incorrect	total
ambiguous	5.1	0.3	0.9		0.9	7.2
ill-formed	2.4	0.3	0.9		1.8	5.4
out of domain	0.6					0.6
reasonable	69.9	4.5	0.9	1.8	9.6	86.7
total	78.0	5.1	2.7	1.8	12.3	

(a) Speech Input

	answer appropriate	answer verbose	error appropriate	error ambiguous	response incorrect	total
ambiguous	5.5	0.3	1.0	0.1	0.7	7.6
ill-formed	2.2	0.1	0.1		1.4	3.8
out of domain	0.6		0.1		0.1	0.8
reasonable	72.1	5.0	1.4	1.0	8.0	87.5
total	80.4	5.4	2.6	1.1	10.2	

(b) Orthographic Input

Table 2: Breakdown of subjective judgements on system responses and user queries for (a) speech input, and (b) orthographic input.

be possible now that systems have become more complex. While some objective criteria exist for individual components, overall system performance should probably incorporate subjective judgements as well.

Thus far, we have not addressed the issue of efficiency, mainly because we have not focussed our attention on that issue. When VOYAGER was first developed, it ran on a Symbolics Lisp machine, and took several minutes to process a sentence. More recently, we have started to use general signal processing boards to derive the auditory-based signal representation, and a Sun workstation to implement the remainder of the SUMMIT recognition system. Currently, the system runs in about 12 times real-time. The approximate breakdown in timing is shown in Table 3. Note that the natural language component and the back end run in well under real-time. Refined algorithms, along with the availability of faster workstations and more powerful signal processing chips should enable the current VOYAGER implementation to run in real-time in the future. On the other hand, the computation is likely to increase dramatically when speech recognition and natural language are fully integrated, since many linguistic hypotheses must be pursued in parallel.

References

[1] Pallett, D. "Benchmark Tests for DARPA Resource Management Database Performance Evaluation," *Proc. ICASSP-89*, pp. 536-539, Glasgow, Scottland, 1989.

[2] Zue, V., Glass, J., Goodine, D., Leung, H., Phillips, M., Polifroni, J., and Seneff, S., "The VOYAGER Speech Understanding System: A Progress Report," These Proceedings.

[3] Zue, V., Daly, N., Glass, J., Leung, H., Phillips, M., Polifroni, J., Seneff, S., and Soclof, M., "The Collection and Preliminary Analysis of a Spontaneous Speech Database," These Proceedings.

Components	Timing (x RT)
Speech Recognition	
Signal Representation	2.5
Phonetic Recognition	4
Lexical Access	5
Natural Language	.2
Back End	.2

Table 3: Breakdown in computation for VOYAGER components.

[4] Zue, V., Glass, J., Phillips, M., and Seneff, S., "The MIT SUMMIT Speech Recognition System: A Progress Report," *Proceedings of the First DARPA Speech and Natural Language Workshop,* pp. 178-189, February, 1989.

[5] Seneff, S., "TINA: A Probabilistic Syntactic Parser for Speech Understanding Systems," *Proceedings of the First DARPA Speech and Natural Language Workshop,* pp. 168-178, February, 1989.