THE COLLECTION AND PRELIMINARY ANALYSIS OF A SPONTANEOUS SPEECH DATABASE*

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

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

ABSTRACT

As part of our effort in developing a spoken language system for interactive problem solving, we recently collected a sizeable amount of speech data. This database is composed of spontaneous sentences which were collected during a simulated human/machine dialogue. Since a computer log of the spoken dialogue was maintained, we were able to ask the subjects to provide read versions of the sentences as well. This paper documents the data collection process, and provides some preliminary analyses of the collected data.

INTRODUCTION

One of the first tasks confronting researchers developing a spoken language system is the collection of data for analysis, system training, and evaluation. Since people do not always say grammatically well-formed sentences during a spoken dialogue with a computer, the currently available *read* speech databases may not capture the acoustic and linguistic variabilities found in goal-directed spontaneous speech. As a first attempt to create a spontaneous speech database[1], we have recently collected a large amount of data from 100 subjects during simulated dialogues with the VOYAGER spoken language system. The purpose of this paper is to document the database construction process, and to provide some preliminary linguistic and acoustic analysis.

VOYAGER is a system that knows about the physical environment of a specific geographical area as well as certain objects inside this area, and can provide assistance on how to get from one location to another within this area. It currently focuses on the geographic area of the city of Cambridge, Massachusetts, between MIT and Harvard University, and can deal with several distinct concepts including directions, distance and time of travel between objects, relationships such as "nearest," and simple properties such as phone numbers or types of food served. VOYAGER also has a limited amount of discourse knowledge which enables it to respond to queries such as: "How do I get *there*?" It can also deal with certain clarification fragments such as: "The bank in Harvard Square." A detailed description of the VOYAGER system can be found elsewhere in these proceedings [1].

VOYAGER is made up of three components. The first component, the SUMMIT speech recognition system [2], converts the speech signal into a set of word hypotheses. The natural language component, TINA [3], provides a linguistic interpretation of the set of words. The parse tree generated by TINA is translated into a query language form, which is used to produce a response. Currently VOYAGER can generate responses in the form of text, graphics, and synthetic speech. The back end is an enhanced version of a direction assistance program developed by Jim Davis of MIT's Media Laboratory [4].

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

[1] We loosely use the term *spontaneous speech* to mean the speech produced by a person "on the fly" when interacting with a computer for problem solving.

DATABASE CONSTRUCTION

We believe that data should be collected under conditions that closely reflect the actual capabilities of the system. As a result, we have chosen to have subjects use the system as if they are trying to obtain actual information. The data were recorded in a simulation mode in which the speech recognition component was excluded. This step was taken partly to avoid long processing delays that would disrupt the human-machine interaction. Instead, an experimenter in a separate room typed in the utterances spoken by the subject, after removing false starts and hesitations. Subsequent processing by the natural language and response generation components was done automatically by the computer.

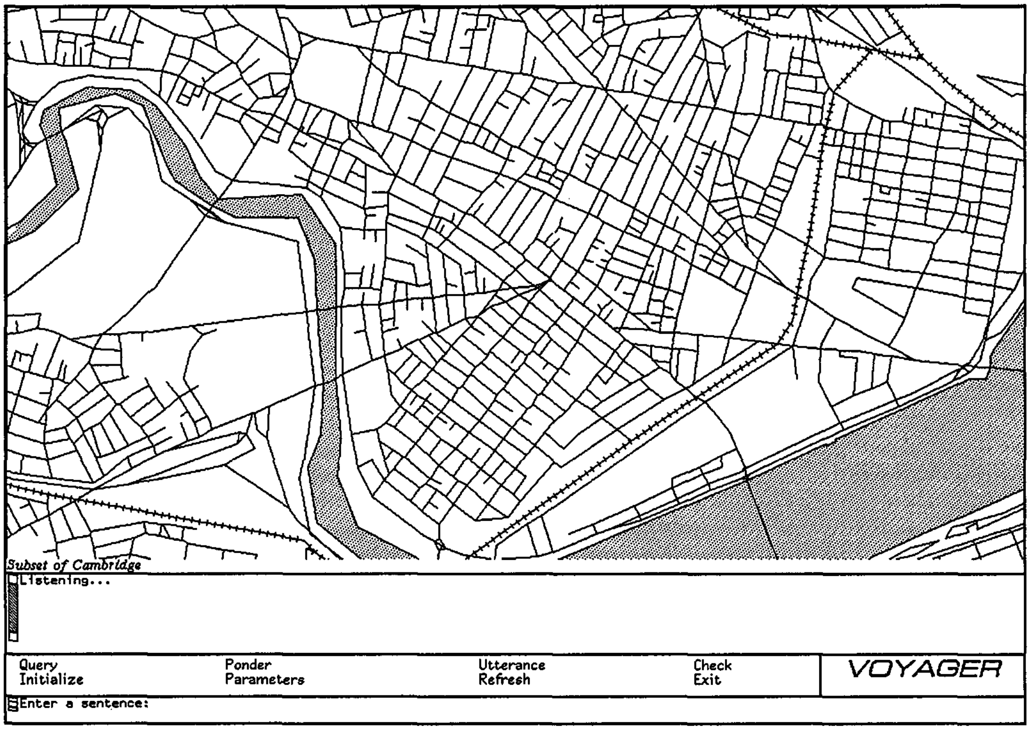

Figure 1: A display used during the simulated data collection process. The type-in window at the bottom was hidden from the subject's view to avoid unnecessary distractions.

DATA COLLECTION

The data were collected in an office environment where the ambient noise was approximately 65 dB SPL, measured on the C scale. A Sennheiser HMD-224 noise-cancelling microphone was used to record the speech. The subject sat in front of a computer console that displayed the geographical area of interest as shown in Figure 1. The console was slaved to the experimenter's console in the adjacent room. The experimenter's typing, shown in the bottom of the display, was hidden from the subject to avoid unnecessary distractions. Two information sheets describing both the knowledge base of VOYAGER and its possible responses were available to the subject. The subjects referred to these sheets from time to time in order to stay within VOYAGER's domain of knowledge.

During a subject's dialogue, both the input speech and the resulting responses were recorded on audio tape. The voice input, minus false starts, hesitations, and filled pauses, was typed verbatim to VOYAGER by an experimenter, and saved automatically in a computer log. The system response was generated automatically from this text, which was also recorded into the log. The system's response typically took a second or two after the text had been entered.

Whenever a sentence contained words or constructs that were unknown to the natural language component, the system would explain to the subject why a response could not be generated. In the event that the queries were outside of the knowledge domain and the system responses could not dislodge the subject from that line of questioning, the experimenter could override the system and trigger a canned response explaining that the system was currently unable to handle that kind of request. Another canned response was available for the case when the subject produced several queries at once.

Each session lasted approximately 30 minutes, and began with a five minute introductory audio tape describing the task. This was followed by a 20 minute dialogue between the subject and VOYAGER. Following the dialogue, the subject was asked to read his or her sentences from the computer log. The resulting database therefore included both a read and a spontaneous version of the same sentence, modulo false starts, hesitations, and filled pauses in the spontaneous version.

Fifty male and fifty female subjects were recruited as subjects from the general vicinity of MIT. They ranged in age from 18 to 59. The only requirement was that they be native speakers of American English with no known speech defects. For their efforts, each subject was given a gift certificate at a popular ice-cream parlor. The entire recording was carried out over a nine-day period in late July. Several of the sessions were also recorded on video tape to document the data collection process.

DIGITIZATION AND TRANSCRIPTION

The recordings made during the data collection were digitized at 16 kHz, after being band-limited at 8 kHz. Special care was used to ensure that false starts, hesitations, mouth clicks, and breath noise were included as part of the digitized utterance. In addition, pré-determined conventions were established, and written instructions provided, for transcribing these non-speech and partial-word events both orthographically and phonetically. We started with the notations suggested by Rudnicky [5], and made modifications to suit our needs.

To date, the entire database of 9,692 utterances has been digitized and orthographically transcribed, including markers for false starts, partial words, and non-word sounds. In addition, an aligned phonetic transcription has been obtained for approximately 20% of the data.

PRELIMINARY ANALYSIS

We have divided the database into three parts according to speakers. Data from 70 arbitrarily selected speakers were designated as the *training* set. Of the remaining speakers, two-thirds were designated as the *development* set, and the rest as the *test* set. In each set, there were equal numbers of male and female speakers. In this section, we will report on the results of some preliminary analysis on parts of this database, carried out over the past few weeks.

GENERAL STATISTICS

From the computer log, we were able to automatically generate some preliminary statistics of the database. Table 1 summarizes some of the relevant statistics for the sum of the training and development sets. Note that the number of sentences refers to the spontaneous ones; the total number collected is double this amount.

As the table reveals, approximately two-thirds of the sentences could be handled by the current version of VOYAGER. The remaining third of the data is evenly divided between sentences with out-of-vocabulary words and sentences for which no parses were generated. These sentences can be used to extend VOYAGER's

Speakers	90
Total Sentences	4361
Avg. Words per Sentence	8.0
Sentences with Action	2854(65%)
Sentences with Unknown Words	740 (17%)
Sentences with No Parse	727(17%)
Sentences with No Action	40(1%)
Words Used	601
Unknown Words	398
Unknown Word Frequency	3%

Table 1: General statistics of the spontaneous speech from the training and development sets of the VOYAGER database.

capabilities. Only a very small amount, about 1%, were parsed but not acted upon. This is a direct result of our conscious decision to constrain the coverage of the natural language component according to the capabilities of the back-end.

The version of VOYAGER used for data collection had a vocabulary of about 300 words which were determined primarily from a small set of sentences that we made up. It is interesting to note that only about 200 of these words were actually used by the subjects. While the number of unknown words appears to be large, they actually account for less than 3% of the total number of words when frequency of usage is considered.

The statistics of this database indicated that an average of slightly less than 50 sentences per subject were collected in each 20 minute dialogue. Thus we believe the database can easily be expanded as the capabilities of the system grow.

ACOUSTIC ANALYSIS

Since time-aligned phonetic transcriptions were already available for part of the database, we performed some comparative acoustic analyses of the spontaneous and read utterances. These preliminary analyses were carried out using slightly over 1,750 sentences from 9 male and 9 female training speakers. While these data represent less than 20% of the recorded data, there were more than 60,000 phonetic events. As a result, the quantitative differences were found to be statistically significant. Rather than exhaustively reporting our findings, we will make a few observations based on some interesting examples.

Figure 2 compares the overall duration of the read and spontaneous utterances. In this and subsequent figures, the thin line denotes read speech, whereas the thick line denotes spontaneous speech. The horizontal bars show the means and standard deviations. These values, together with the sample size, are also displayed to the right. The figure suggests that spontaneous utterances are longer than their read counterparts by more than one-third. However, there is much more variability in the duration of spontaneous speech, as evidenced by its considerably larger standard deviation.

There were nearly 1,000 pauses found in the spontaneous sentences in our dataset, or more than one per sentence on the average.[2] In contrast, there were only about 200 pauses found in the read sentences. As Figure 3 reveals, the pauses in spontaneous speech are about 2.5 times longer on the average than those in read speech. Their durations are also much more variable.

There are nearly 400 non-speech vocalizations found in this database, including mouth clicks, breath

[2] We make a distinction between pauses, epenthetic silences and stop closures. Only those silence regions that do not have phonetic significance are labeled as pauses.

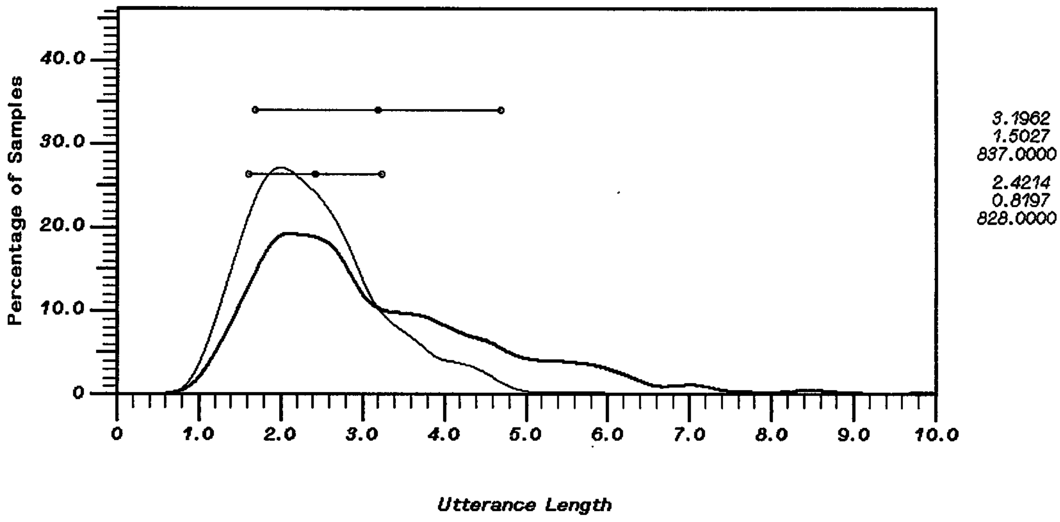

Figure 2: Normalized histogram of overall duration for read (in thin lines) and spontaneous (in thick lines) utterances for 9 male and 9 female speakers.

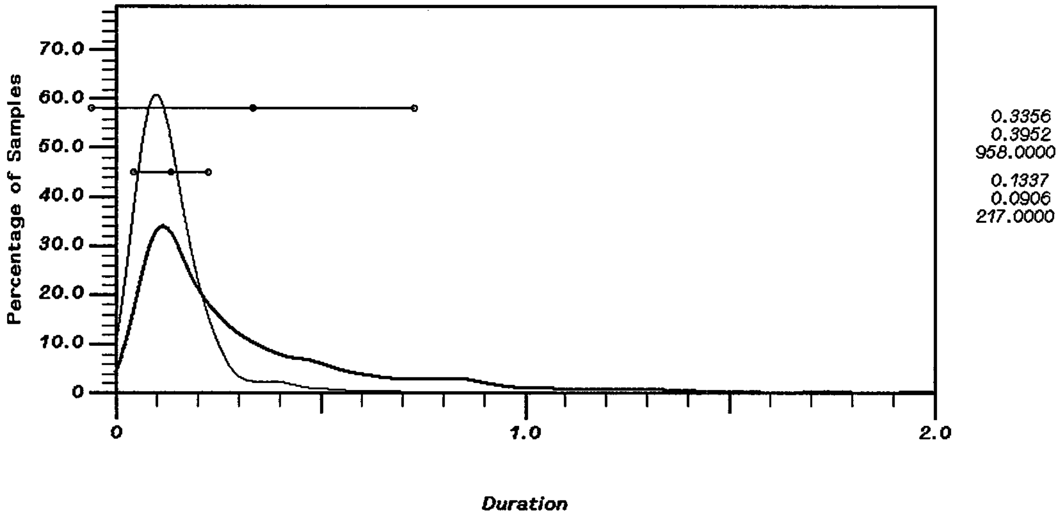

Figure 3: Normalized histogram of pause duration for read (in thin lines) and spontaneous (in thick lines) utterances for 9 male and 9 female speakers.

Category	Read Speech	Spontaneous Speech
TOTAL	103	269
Mouth Clicks	60	106
Breath Noise	37	117
Filled Pauses	0	30
Others	6	16

Table 2: Number of occurrences of non-speech vocalizations for read and spontaneous speech from 9 male and 9 female speakers.

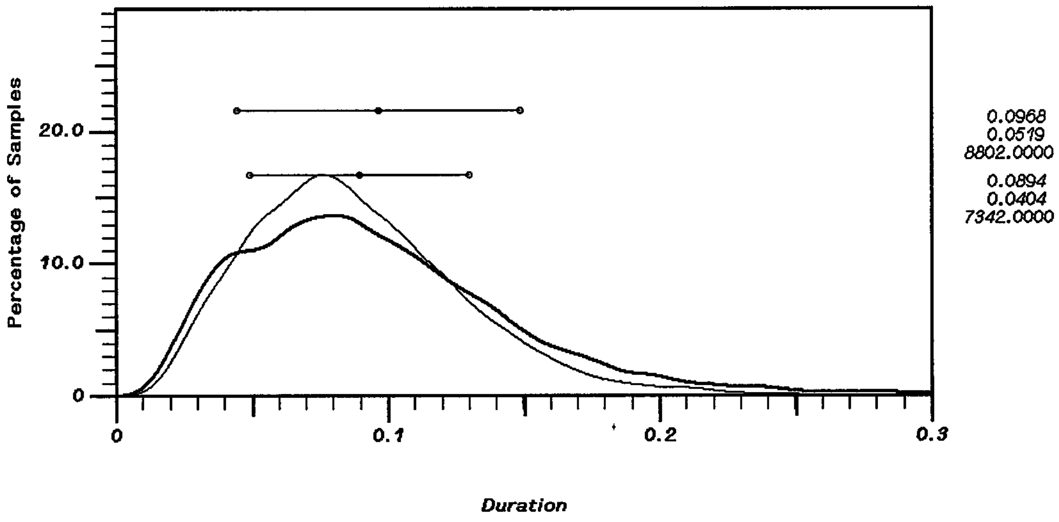

Figure 4: Normalized histogram of vowel duration for read (in thin lines) and spontaneous (in thick lines) utterances for 9 male and 9 female speakers.

noise (both inhaling and exhaling), and filled pauses such as "um, "uh," or "ah." Their distributions are shown in Table 2. Non-speech vocalizations occur about 2.7 times more often in spontaneous speech than in read speech. Almost all of the clicks appear at the beginning of sentences for read speech, whereas 25% of them occur sentence internally in spontaneous speech. Similarly, more than 20% of the breath noise occurs sentence internally, with five times as many in spontaneous speech as in read speech. All the filled pauses occur in spontaneous speech, two-thirds of them sentence internally.

When we measured the durations of individual phonemes, we found very little difference between the two speech styles. Figure 4, for example, shows that the average vowel durations for read and spontaneous speech are 89 ms and 95 ms, respectively. Occasionally, we observed unusually long vowels in the spontaneous speech. They almost always correspond to words like "is" or "to," when the subject tries to decide what to say next. An example is shown in Figure 5.

LINGUISTIC ANALYSIS

When the database was transcribed orthographically, false starts and non-words such as "ah," "um," or laughter were explicitly marked in the orthography. Therefore, it is possible to perform a statistical analysis of

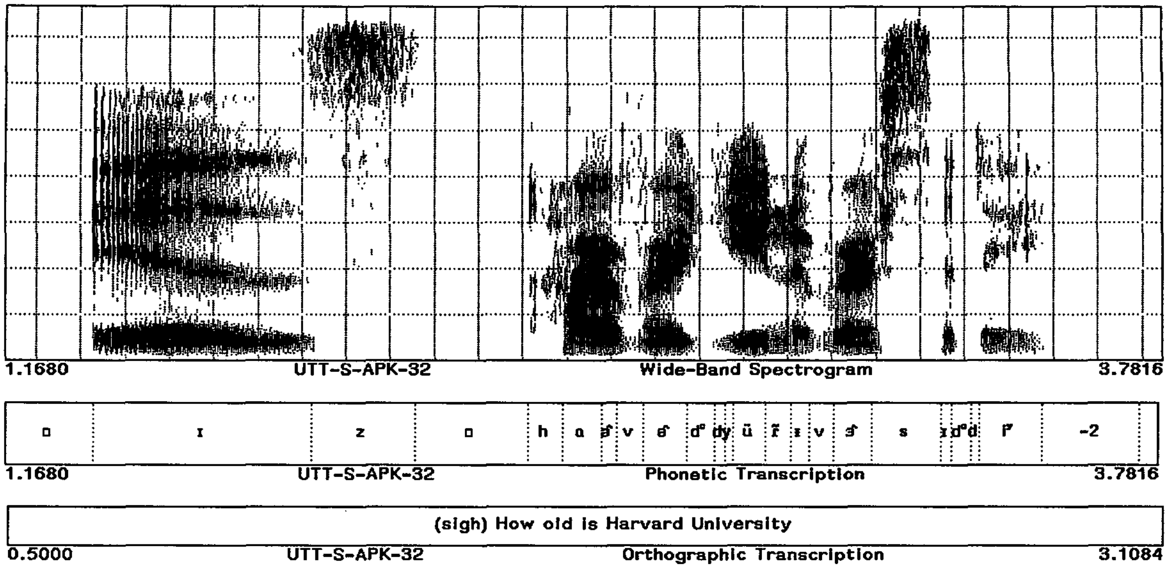

Figure 5: Spectrogram of the sentence, " (How old) is Harvard University?" showing a lengthened "is".

how often such events occurred. We distinguished between non-words internal to the sentence and non-words at the beginning or end of the sentence. In the training set containing about 3,300 spontaneous sentences, more than 10% of the sentences contained at least one of these effects. An additional 25% contained mouth clicks or breath noise, which may be a less serious effect. About half of the non-words appear sentence internally.

False starts occurred in almost 4% of the spontaneous sentences. Table 3 categorizes the words following false starts in terms of whether a given word was the same as the intended word, a different word in the same category, a new linguistic category, or a back up to repeat words already uttered. An example is given for each case. Over 40% of the time, the talker repeated words after the false start. In order to recognize such sentences correctly, the system would have to detect false starts and back up to an appropriate earlier syntactic boundary.

Category	%	Example
Same Word	32	Wh(at) what's the street address of Cajun Yankee?
Same Category	8	Show me the intersection of Har(vard) Hampshire Street and Cambridge Street.
New Category	19	Where is the nearest restaurant to Memori(al) 305 Memorial Drive?
Back Up	41	How do I get from the Ma(rriott) from the Marriott to Bel Canto's?

Table 3: Breakdown of false starts in training sentences.

We have examined the linguistic content of the training sentences and have begun to use them to expand VOYAGER's coverage. We have found a number of new linguistic patterns that are entirely appropriate for the domain, as well as a few recurring concepts currently outside of the domain that would be reasonable to add. An example of the latter is a comparison between two objects, such as "Which is closer to here, MIT or Harvard University?" In addition, we were surprised at the number of ways people said certain

Your directions were not good.
Do you like their ice cream?
Does the *Baybank* serve ice cream?
Where is my dog?
Does Hampshire and Broadway intersect?
How far am I from Central Square *to Cajun Yankee*?

Table 4: Examples of problematic sentences from the database.

things. For instance, for sentences like "What is the phone number *of* MIT?" the prepositions, "of," "at," "for," and "to" were all used. In a similar vein, the "from location" in a sentence requesting distance between two objects occurred in five distinct syntactic locations within the sentence. Users also sometimes spoke ungrammatically, violating both syntactic and semantic constraints. In other cases they asked abstract questions or questions involving judgment calls that would be inappropriate for VOYAGER to handle. Some of these sentences were probably uttered due to curiosity about how the system would respond. Some examples are given in Table 4.

SUMMARY

This paper documents our initial effort in developing a spontaneous speech database, and reports some preliminary analyses of the collected data. We found the process of data collection to be relatively straightforward, and we believe it will be fairly easy to collect more data at a later stage, after VOYAGER's capabilities have improved. In fact, we believe that incremental data collection done this way can be quite effective for development of spoken language systems.

Our preliminary analysis of these data has already indicated some significant differences between spontaneous and read speech. We are also beginning to use the database to train and evaluate the VOYAGER system, both at the acoustic-phonetic and the linguistic levels. The results of the preliminary evaluation are described in a companion paper [6].

Acknowledgments

We gratefully acknowledge the assistance of other members of the Spoken Language Systems Group for helping us with data collection and transcription. We would also like to thank our subjects, who offered their time and effort for the sake of science and ice cream.

References

[1] 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.

[2] 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.

[3] 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.

[4] Davis, J.R. and Trobaugh, T.F., "Directional Assistance," Technical Report 1, MIT Media Laboratory Speech Group, December 1987.

[5] Rudnicky, A.I. and Sakamoto, M.H., "Transcription Conventions for Spoken Language Research," CMU School of Computer Science Technical Report CMU-CS-89-194, 1989.

[6] Zue, V., Glass, J., Goodine, D., Leung, H., Phillips, M., Polifroni, J., and Seneff, S., "Preliminary Evaluation of the VOYAGER Spoken Language System," These Proceedings.