Artificial Intelligence and Language Processing

Bonnie Webber Editor

Natural Language with Discrete Speech as a Mode for Human-to-Machine Communication

ALAN W. BIERMANN, ROBERT D. RODMAN, DAVID C. RUBIN, and J. FRANCIS HEIDLAGE

ABSTRACT: A voice interactive natural language system, which allows users to solve problems with spoken English commands, has been constructed. The system utilizes a commercially available discrete speech recognizer which requires that each word be followed by approximately a 300 millisecond pause. In a test of the system, subjects were able to learn its use after about two hours of training. The system correctly processed about 77 percent of the over 6000 input sentences spoken in problem-solving sessions. Subjects spoke at the rate of about three sentences per minute and were able to effectively use the system to complete the given tasks. Subjects found the system relatively easy to learn and use, and gave a generally positive report of their experience.

1. INTRODUCTION

The advent of a number of commercial discrete speech recognition systems[1] and their use in industrial applications [27] raises the question of whether or not such machines may be useful for input to natural language processors. Specifically, one can ask

a) how might a natural language processor perform in conjunction with such input devices, and

b) how habitable would the resulting voice-interactive systems be for real users?

To study these issues, the Voice Natural Language Computer system was constructed [6]. This system allows a user to display one or more matrices on a computer terminal and to manipulate them with spoken English imperative sentences. If an undesired behavior is observed, the user may request a "backup" and rephrase the command appropriately.

The voice recognition device used is a Nippon Electric Corporation DP-200 Discrete/Connected Speech Recognizer which is a speaker-dependent machine that requires a person to register in advance his or her pronunciation of each vocabulary word one or more times in a training session. This recognizer operates in the "discrete speech" mode, which requires a user to pause for about 300 milliseconds after each word, and the "connected speech" mode, which allows word bound-

[1] For example, numerous such processors are described in the 1984 issues of *Speech Technology*, Media Dimensions, New York, N.Y.

aries to merge as long as individual word pronunciations are clear. Our work investigates the use of connected speech [14] and discrete speech, which is reported here.

Discrete speech recognition is far more error-free than connected speech recognition because the processor does not have to guess at word boundaries. Thus, larger vocabularies, and hence richer domains of discourse, are achievable with discrete speech at a cost of lower speech rates and some user inconvenience. In many applications where speech is the preferred mode of input, these disadvantages are far outweighed by the advantages of robustness and greater expressive power. Furthermore, users seem to be able to learn to speak machine recognizable discrete speech more easily than they are able to learn to speak machine recognizable connected speech. A more detailed comparison of the two modes is found in [11, 17, 29].

In the system under discussion, the voice recognizer drives an error-correcting parser as described in the next section. The parser sends output to a language semantics processing system as described in [4–6]. Then, a domain-simulation module executes the desired action and displays the result to the user. The total system design will eventually include a touch-sensitive screen to allow users to point to objects as they speak, and a voice response system for prompting and error messages. This article, however, will examine only issues related to voice recognition.

Evaluating speech recognition systems is a difficult undertaking because so many factors affect the variables to be measured. Lea [24] mentions over 80 factors which affect performance, including vocabulary size, vocabulary confusability, the physical and emotional state of the user, adjustments to system parameters, the type and placement of the microphone, and environmental noise. Error rates, speed of entry, and user satisfaction may all vary drastically on the same system under different test conditions. To place our experiment in the proper perspective, we have classified voice test conditions into the following five categories in order of increasing difficulty:

A) Measurements are made in the manufacturer's laboratory reading lists of words under optimal conditions.

B) Measurements are made reading lists of words in our laboratory environment.

C) Measurements are made reading sentences in our laboratory.

D) Measurements are made as a user utters commands to our system in a problem-solving situation in our laboratory.

E) Measurements are made as a user utters commands to a system in a problem-solving situation in the user's own work environment.

Under Condition A, 99 percent word recognition rates are usually reported. Under Condition B, the clarity of the speaker, microphone placement, and other factors may vary so that recognition may fall several percent. Under Condition C, the reader will unconsciously inflect words placing stress at key points in the sentence and allowing falling inflection near the end. This will further reduce recognition rates. Under Condition D, the user will stop thinking about how to speak to the machine and will begin concentrating on problem-solving issues. Moreover, the user may speak ill-formed sentences or use disallowed vocabulary. All of these effects introduce additional errors with a concomitant slowing of entry speed. In the final condition, numerous other environmental factors are bound to further distract the user from speaking in the carefully paced style that can be recognized by a machine. Although we feel that Condition E is of highest interest, circumstances confine us to Condition D, which may or may not be an approximation to Condition E, depending on the details of the laboratory environment. Our laboratory was, for the most part, free of noise that might affect recognition, and test subjects were usually allowed to proceed uninterrupted. Otherwise, the environment was similar to what might be found in an ordinary office.

In our experimental study we were interested in the following kinds of questions:

1. Learnability: What training would be required to teach users to speak in machine-recognizable sentences? What additional learning would occur while individuals used the system and how fast would it occur?

2. Correctness: What word error rates would be delivered by the voice recognizer and to what extent would these errors be correctable by a machine? What sentence error rates would result? Would users be able to successfully complete tasks?

3. Timing: How fast would users speak individual commands and how many commands would be given per minute? How fast could tasks be completed?

4. User Response: How would users feel about speaking machine-recognizable commands? How would they judge their ability to do useful work in such a manner?

Toward the goal of at least partially answering these questions, an experiment was run in which paid subjects used our Voice-driven Natural Language Computer (VNLC) system with the DP-200 recognizer to solve a series of problems using discrete speech. Following a two-hour training session, each subject used the system to solve simple problems for a total of four hours. Two subjects were retained for an additional six hours of testing to examine longer-term usage of the system. Over 6000 utterances were spoken by the subjects in these sessions. This article reports the major findings related to the above questions. In addition, one

expert speaker with extensive experience and considerable knowledge of the system itself solved the same problem set as the novice subjects. In the following sections, some practical considerations concerning voice recognition systems will be given and then the experiments and their results will be described. Finally, this work will be compared to other projects in speech understanding systems.

2. VOICE RECOGNITION PRAGMATICS

The most widely used voice recognition systems are based on the pattern matching paradigm [20, 33]. Initially, the user registers one or more samples of every vocabulary word in the machine; at later times, recognition is done by comparing the time-spectral characteristics of unknown words with the known samples. The unknown word is assumed to be identical to that stored prototype which is closest to it by some distance measure. But if the unknown word has no near matches, the recognizer may reject it by refusing to make a selection. Connected speech on such machines is typically done by a two-level dynamic programming algorithm that guesses word boundaries at one level and matches words at the other level.

Such recognizers have been used in many industrial situations for voice data entry to computers [27]. For example, in package sorting or assembly line inspection applications, workers may vocalize digit sequences or other information to the machine while carrying on other tasks with their hands. However, there is very little mention in the literature of the use of these recognizers with natural language processors where vocabularies are higher and the information transmitted to the machine more complex. Because the slow mechanical quality of machine recognizable complete sentences may be bothersome to users, and the inflections of fully formed sentences may reduce recognition accuracy, potential users have shied away from installing such systems. One of the purposes of our work is to evaluate systems of this type. An example of such use outside of our laboratory is the "Put that there" system at the MIT Architecture Machine Project [34].

Four kinds of error may occur during recognition of discrete speech: rejection of legal input, substitution of the wrong word for the word spoken, false acceptance of illegal input, and failure to respond to legal input. *Rejection* occurs when the speech recognizer cannot find a sufficiently close match between the spectral pattern of the input and any of the spectral patterns of the prestored reference words. Rejection of input is the correct response if the input is a nonvocabulary word, or a nonword altogether, such as breath noise. However, if the input is a legal vocabulary word, this is a *rejection error.* A *substitution error* occurs when the speech recognizer reports a word other than the one spoken. This happens when the spectral pattern of the input is matched to the pattern of the wrong reference word.

Substitution errors tend to follow predictable patterns and, therefore, are often automatically correctable. In order to detect and possibly correct such errors, the VNLC language processor receives a set of one or more word guesses in each word position in the input sentence. The first guess comes from the DP 200's best match and the additional guesses come from an historical record of known word confusions. For example, if the recognizer commonly identifies the pronunciation of "divide" as "five," each recognition of "five" will also include "divide" as an alternative. After incompatible adjacent pairs have been eliminated by the scanner preprocessor, the sequence of word sets are then passed on to the parser and the semantics processor which attempt to select a single word for each slot such that the sequence is a meaningful English command. The first such sequence found is executed for the user who can view the result on the screen and judge its correctness. If no legitimate English command can be found, the processor returns a prompt to the user to "please rephrase the request."

False acceptance errors occur when a nonvocabulary word or a noise such as a cough is reported to be a word in the vocabulary. *Failure to respond to legal input,* the fourth type of error, is usually the result of faulty gain setting or very soft-spoken input. This was not a problem during this experiment. The frequency of all the above errors will be affected by certain internal parameter settings of the recognizer. For our study, these were set to levels recommended by the manufacturer and were not altered during the experiment.

3. THE EXPERIMENTAL PROCEDURE

Nine volunteers from a college mathematics course, who ranked themselves above average in mathematical ability, were selected as subjects. Their first session of about 60 minutes was used to establish their reference templates of the 100 vocabulary words. Each word was registered once using rising inflection (as in "Did you say 'multiply'?"), once using flat inflection (as in "The word 'multiply' is misspelled."), and once using falling inflection (as in "Now speak the word 'multiply'."). Forty-five difficult-to-recognize words were registered three additional times using the same inflection patterns for a total of 435 word samples. The subject was then asked to read the vocabulary list back to the machine in order to detect recognition problems and some words were reentered if necessary.

In the second session, the subject was seated behind a computer display terminal with a head-mounted microphone and introduced to the voice natural language processor in a 30–60 minute training session. The experimenter read a tutorial description of the basic system capabilities and at appropriate times requested that the subject speak specific commands to illustrate the facilities being described. As each input word was spoken, the recognizer either returned a low-pitched audio beep to the headset to indicate a rejection or displayed

its best guess of the word on the screen. In case of a substitution error, the subject could say "correction" and repeat part or all of the sentence, or the subject could ignore the displayed words and depend on system error correction to find the desired meaning of the command. Subjects were not initially informed about all the error-correcting abilities of the system, but rather were allowed to discover them as the experiment progressed.

Subjects were required to end each utterance with the word "over" as a request for command execution. The word "forgetit" could be spoken at any point to terminate an utterance without command execution and prepare the machine for a new command. Utterances ending with either "over" or "forgetit" are called *transactions* in this article. Transactions ending with "over" that were processed by the system are called *successful transactions*. Transactions ending with "forgetit" and transactions which resulted in error messages or no response are called *unsuccessful transactions*. The *transaction time* is the total elapsed time from the beginning of one transaction to the beginning of the next, and includes speaking time, processing time, and any user delays before beginning the next utterance.

The experimental task to be performed was the solution of a set of three simultaneous linear equations in three unknowns. In each case, the subject would use spoken commands to display the coefficients in a three-by-four matrix. The method of solution was left to the subject but usually involved obtaining an identity matrix in the first three columns and the solutions in the fourth. Subjects were not taught the use of looping or procedural capabilities [15] so their solution included only "straight line" code. The advantage of this kind of problem is that it is easy for the subject to comprehend what to do and it evokes a significant number of diverse vocal commands. Subjects solved problems at their own rate and were paid when they completed the experiment. The experimental sessions were tape recorded and the machine's recognition of each command was retained by the computer. After the experiment, each subject filled out an interview form evaluating his or her experience.

One of the authors has had considerable experience with the system, and may be considered an "expert speaker." For purposes of comparison, he solved problems under the same experimental conditions as the other subjects, and the data were subjected to the same analysis.

4. RESULTS

During problem solving, each subject uttered at least 400 transactions. Analysis was conducted as follows: The first 15 transactions were scored, 60 were skipped, the next 15 scored, and so on until at least 6 *intervals* of 15 transactions each were analyzed. Three kinds of data were collected, related to the processing of words, the processing of transactions, and the completion of tasks.

TABLE I. Summary of Data per 15-Transaction Interval. 6 Intervals: 9 Subjects

Parameter	Mean	Low	High	Expert
Number of words	116.0	93.8	133.8	94.2
Words per minute	46.5	40.8	57.0	79.4
Transaction time (seconds)	19.4	14.8	24.9	8.9
Number of successful transactions	11.2	9.5	14.3	13.5
Number of rejection errors	4.9	.8	12.5	1.3
Number of correctable substitution errors (c-subs)	2.8	.2	5.2	2.0
Number of incorrectable substitution errors (i-subs)	5.5	.8	10.0	.8
Number of false acceptance errors	.6	.0	1.3	.0
Number of total errors	13.1	2.0	25.5	4.1

TABLE II. Data From the Tutorial Phase (Normalized to 15-Transaction Interval) 9 Subjects

Parameter	Mean	Low	High
Number of words	109.4	82.0	133.0
Number of successful transactions	11.7	7.0	13.0
Number of correctable substitution errors	2.8	.0	6.0
Number of incorrectable substitution errors	9.9	3.1	19.6

Table I presents the results for nine parameters over the six intervals that all nine subjects completed. The data are presented as mean values over all subjects, the minimum value and the maximum value. The latter two show considerable variation that occasionally exceeded a tenfold range between best and worst performance. The last column contains the results obtained from our expert speaker.

Several parameters exhibited a trend with time although these effects were not statistically significant. For example, the speaking rate in words per minute increased from 42.3 in the first interval to 50.3 in the sixth. Correspondingly, the transaction time decreased from 23.2 to 15.9 seconds over the first six intervals. None of the error measuring parameters (the last five rows of the table) showed a significant trend over the intervals measured.

Since the error rates did not decline significantly, as we had expected, we decided to analyze the data recorded during the tutorial sessions. Tape recordings were not made during this phase of the experiment so only four data items were available. These results, shown in Table II, were normalized to the 15 transactions per interval to allow direct comparison with Table I. Although the rate of *i*-subs was greater during the tutorial, the other three parameters are remarkably similar to the data shown in Table I.

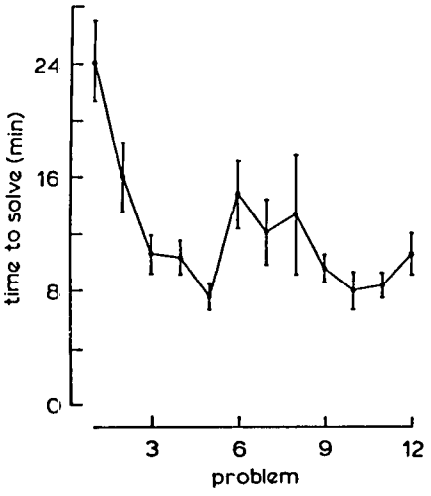

FIGURE 1. Problem Solution Time Versus Problem Number

The time taken by the subjects to solve a given problem was also recorded. As shown in Figure 1, the subjects required an average of 24 minutes to solve the first problem attempted, and from 7 to 15 minutes for subsequent problems. Analysis of variance showed that the mean times to solve problems are significantly different at $P < .01$.The least significant difference criterion [8] applied to the ordered list of means indicated that the largest mean (i.e., that of Problem 1) was significantly different from the second largest (that of Problem 2), but that the second and remaining means did not differ from one another at the $P < .05$ level. (It was found in an earlier experiment [5], that first year programming students, using a version of this system with typed input, required about 34 minutes to solve a set of three linear equations.)

A breakdown of unsuccessful transitions into these various causes is given in Table III. Of the 1365 transactions actually scored, (about one-fifth of the total), 1053, or 77.1 percent were successful, and 312 or 22.9 percent were unsuccessful. Most of the unsuccessful transactions were in the category of simple "forgetits." These were utterances in which the subject determined by inspection that the sentence was incorrect and terminated the transaction rather than attempting to correct the errors. The next largest category of failed transactions was "user error." This consisted primarily of errors in use at the correction facility: either too many or too few words were deleted so that the resulting utterance could not be correctly parsed. Failures marked as "i-sub errors" were those in which the speech recognizer made an incorrectable substitution which was not noticed by the subject, but which caused the input sentence to be immediately rejected as syntactically impossible by the scanner-preprocessor. In 23 transactions, subjects spoke what appeared to be a well-informed, syntactically correct sentence, to which the computer did not respond correctly. These are listed as "system errors." In a few cases, the subject had a change of mind after speaking a well-formed sentence (e.g., the subject meant to say "Double row two and . . ." rather than "Double row one and . . .") and simply terminated the transaction. Some subjects attempted unimplemented variations of operations, such as "Add three times row two to row one," rather than the acceptable "Add the product of three and row two to row one." Logouts occurred when a subject became impatient with a time-consuming parse and terminated the process.

Data on subject responses to a questionnaire are shown in Table IV. Six of the questions required a numerical response, and the mean results are shown. The subjects appear to have enjoyed learning the system, and found it easy to learn (Questions 1 and 2). Although enjoyment and ease of use (Questions 3 and 4) did not rate as highly as ease of learning, both are scored favorably. Ease of use (Question 4) would have a higher rating were it not for one user who scored this question 1, the lowest rating. The worst score assigned this question by any other subject was 4, the neutral rating. Overall, most subjects rated the system as not tiring to use (Question 5) although there was some disagreement on this point. The subjects strongly preferred the voice-driven computer system to working with pencil and paper, and they somewhat preferred it to using typed input (Questions 6a and 6b). However, several subjects expressed a preference for a preprogrammed algorithm which required only that the data be entered into the computer, and which would then automatically calcu-

TABLE III. Breakdown of Unsuccessful Transactions (All Subjects, All Scored Transactions)

Category	Number	Percent of total	Percent of failures
Forgetit	184	13.48	58.97
User error	43	3.15	13.78
i-sub errors	28	2.05	8.97
System error	23	1.68	7.37
Change of mind	17	1.25	5.45
Unimplemented	9	.66	2.88
Logout	8	.59	2.56

TABLE IV. Subject Response to Questionnaires

Question	Mean	High	Low
1. Enjoyed learning system	6.7	7	6
2. Found learning easy	6.3	7	5
3. Enjoyed using system	6.3	7	4
4. Found use easy	5.0	7	1
5. Found system tiring	3.1	5	1
Prefer VNLC to:			
6a. Pencil and paper	6.4	7	5
6b. Using typed input	5.7	7	3

7 = highest degree of agreement
1 = highest degree of disagreement.

late the results. The subjects gave varied responses when queried about their likes and dislikes with regard to the voice-driven system. In general, they enjoyed the novelty of being able to talk to a computer. Several subjects expressly liked the ability to use pronoun referents to objects such as rows and entries, and the ability to use conjunctions to specify more than one operation within a single sentence. Dislikes included the necessity to pause between words (i.e., discrete rather than connected speech) and the long period sometimes required for the parser to fail on a syntactically-incorrect sentence. Related to the dislikes were suggestions which included implementing connected speech, and putting a time limit on the parsing phase. Several subjects suggested increasing the vocabulary size.

5. DISCUSSION

The four questions presented in the introduction will be discussed here.

Learnability: What training was required and what additional learning occurred during system use? The amount of formal training was adequate for the subjects to use the system since all of them satisfactorily solved all problems. After training, two noticeably different kinds of learning were observed during the experimental session. The first concerned the use of the natural language processor in solving this particular type of problem. In Figure 1, one can see that the first one or two problems were solved rather slowly but all after that required roughly the same amount of time. The learning transient was less than one hour.

The second kind of learning, the acquisition of machine recognizable vocal skills, has both a very fast and a very slow component. By the time the tutorial was underway, the subjects had mastered the system to a remarkable degree, as shown in Tables I and II. Progress beyond that point, however, was much slower. For example, the various error rates of the test subjects, (Table I), showed only barely perceptable improvement over the course of the experiment. However, our experienced speaker, also shown in Table I, had an error rate one-third that of the subjects' mean, although he spoke approximately twice as fast. This suggests that refining voice skills is a long-term process.

These results are consistent with our experience. In a previous experiment using only typed input [5], we found that about one hour of training was adequate for most subjects to learn the basics, if not the subtle aspects, of the natural language command system. We also felt that with motivated users voice input could be used effectively with relatively little training (under two hours). The slow rate at which subjects progressed towards their apparent ultimate abilities for vocal machine communication was unexpected. We had hoped that at least the error rates of our longevity subjects (those who spent 10 hours with the system) would improve greatly, but this was not the case. None of our subjects reached the skills of our highly experienced speaker, although there was no reason to suppose that some of them would not have approached or surpassed this level of performance given sufficient time. Apparently, the learning curve is quite flat as users approach their ultimate skill level.

Correctness: What were the error rates and how much could be corrected automatically? The average word error rates (Table I) were in the range of 8–12 percent as can be expected in experimental situations of category (D). These varied widely across subjects as has often been observed in the literature. (See, e.g., [3, 24, 30].) Substitution errors averaged 7.1 percent corrected to 4.7 percent by the system. About one-fourth of all transactions failed to be executed as desired and had to be rephrased. More extensive error correction methods are needed as described, for example, by Fink [14].

The sentence error rate is similar to that observed in experiments on various typed input systems such as LADDER[28], TQA[10], and NLC[5]. Typed input systems usually have much larger vocabularies and broader grammars than VNLC and most of their failed transactions come from users pushing vocabulary or syntax beyond the implemented capabilities. With VNLC, the limited vocabulary resulted in simple grammatical constructions which seldom failed and errors were due primarily to voice misrecognition.

Although the error rate tended to slow the speed at which transactions could be entered successfully and also frustrated and tired our subjects more than an error-free system would, it did not prevent them from carrying out the designated tasks. Moreover, the errors almost never caused the system to carry out a wrong action, that is, one not intended by the user. Rather, in nearly every case, errors prevented the system from carrying out any action, although theoretically wrong actions were quite possible.

In a recent series of tests, we observed substantially fewer *user-induced* errors with discrete speech than with connected speech. With connected speech, inexperienced users often forget to speak in the careful manner required by the recognizer and lapse into a style where rapid bursts and slurred words occur frequently. The regimentation of using discrete speech seems to discourage this behavior, although highly experienced speakers avoid this problem and can achieve low error rates in either mode.

Timing: How fast did subjects speak commands and complete tasks? Various experiments have reported discrete speech word rates between 20 and 50 words per minute, though rates over 40 are considered "burst" rates. One author [29] claims nonburst speeds of over 30 are impossible, although he assumed a simultaneous tasking situation which we did not provide. Speaking sentences in the manner of this experiment constitutes a burst style of speech, which may explain why rates of 40–80 words per minute were achieved.

Concerning sentence rates, subjects entered transactions at the rate of over three per minute with an apparent trend toward faster speech after several hours of

TABLE V. Vocabulary Sizes

Source	Calculation method	Number of types for 95 percent	Comments
Written			
Lorge's magazine count	Calculations from Carroll [7]	11,000	Sample of 5 million words from popular magazines between 1927 and 1938.
Kucera and Francis [22]	Calculation from Carroll [7]	16,500	Diverse sample of 5 million words of written English, where a word is any string of characters separated by spaces.
Spoken			
Howes' spoken count [19]	Read from Table	2,550	250,000 words of monologue.
Dahl's spoken American English [9]	Read from Table	2,243	About one million words of discourse from recorded psychoanalytic cases.
French, Carter, and Koenig [16]	Graphic	600	Words from telephone conversations Names, titles, exclamations, numbers "uh," "er," etc., were omitted as were interjections such as hello, good-bye, all right, so count may be high or low.
Howes, personal communication	Read from table	237	Control tower.
Work done by this project	Hand calculated	435	Vocabulary of weather reports and forecasts, 8265 words.

using the system. Natural language input to computers has not previously achieved this rate in such an experimental situation. Typed input systems have been observed to be used at rates of about one transaction per minute [5]. Typical conversational English is at the rate of about 5–15 sentences per minute.

User Response: How did subjects feel about speaking machine-recognizable commands? Perhaps the most critical factors for the successful practical use of voice input are user attitude and motivation. There are many reasons for an individual *not* to be positively disposed towards a voice system. It may be perceived as hard or impossible to use; it may appear to have the potential of causing co-workers to be displaced; it may provoke unnamed anxiety and hostility merely because it is, in some sense, "a computer" [26]. Our subjects, students from a college math class, do not necessarily represent the kind of person likely to use a voice system routinely. They were an ideal group of subjects for this experiment. Had a significant number of them balked at the system, it would be an ill omen indeed for the practical use of voice. As it turned out no subject displayed any serious hostility toward the system, and every subject who began the experiment completed the allotted hours of work, and did so willingly and cooperatively. There is, then, a segment of the population who can use voice input successfully with very little training and little motivation other than curiosity and the low pay students on a university campus receive. Whether these results would apply to other populations such as workers on an industrial shop floor is a question for future studies.

Other Comments: Vocabulary Size. Our project is concerned with the question of how many words are sufficient for natural language in a restricted domain of discourse. Obviously, limitation to a vocabulary of 100 words stretches the concept of "natural language." Our working definition of natural language requires that it include a broad enough vocabulary and syntax so that the user can say what he or she wants with only a minor effort to adapt the manner of speaking to the listener. Some adaptation is, however, normal and occurs in everyday conversations.

One can gain a rough idea of vocabulary sizes for various environments by examining the literature. Table V gives the numbers of words needed to account for 95 percent of the words for discourse in various studies. The other five percent of words used can involve a nearly unbounded vocabulary in such environments because of the introduction of uncommon words, proper names, and so forth. The seventh entry in Table V is the result of our tabulation of the vocabulary used to report and forecast local and national weather as broadcast over a radio channel dedicated to that purpose.

It appears that while a vocabulary of 100 words was satisfactory for the controlled experiment described here, it is, in general, not adequate for typical applications. Much larger vocabulary capabilities for machines may become available in the near future [21, 35, 38].

For example, Kaneko and Dixon [21] have demonstrated a two-stage recognition procedure which can currently handle a vocabulary of 2000 words in "near real time" with a word recognition rate of 86.5–94.5 percent. Our results show that a word recognition rate in this range can lead to reasonably satisfactory natural language performance. If such recognition systems can be perfected to the point that they can be integrated into the style of processor we have built, a wide variety of potential applications will be within reach.

Clearly, much more precise information is needed on the specific characteristics of natural language in varying environments including vocabulary sizes and the types of grammatical constructions needed. If a categorization of tasks were available specifying the language required in each case, it would be possible to match a given system's capabilities to the tasks that it can successfully cover. Our project prefers to work with task-oriented environments, such as occur with matrix or text manipulation systems, rather than database retrieval environments which are much more typical in natural language research. See, for example, [10, 13, 18, 31, 37, 39, 40]. With task-oriented environments, there is an initial state, goals, and sequential progress toward those goals resulting in dialogues with more form and internal cohesiveness. Furthermore, vocabularies are smaller, execution time can be shorter, and correctness is less of a problem because of the visual feedback.

6. CONCLUSION

The work reported in this article has achieved two results. First, the measurements indicate that current discrete word recognition machines are of adequate quality to enable the construction of small vocabulary (100 word) natural language interactive systems. Such systems can be capable of recognizing most (77 percent) of the sentences spoken by cooperative, intelligent users in the real-time accomplishment of tasks (Category D test) after relatively little (under two hours) training. These conclusions are based on observation of nine subjects speaking over 6000 sentences. Most problems arise from errors delivered by the recognizer and a challenging research area addresses the automatic correction of such errors.

Other data points in speech understanding systems have been established by other projects. For example, the Hearsay-II system [13] processed connected speech with a larger vocabulary (1011 words) and achieved a sentence recognition rate of 91 percent. This result was achieved in a nonreal-time environment with one subject speaking 22 sentences in a Category C testing environment. This system was tuned to the individual speaker by a process of generalizing from speaker-specific training data which was manually labeled by a speech-processing expert [13 (p. 224)]. Similar results were also achieved by the HARPY system [25] in a test where five subjects spoke 184 sentences. Other speech understanding systems [2, 32, 39, 40] have been developed to process connected speech and are described in various standard references, such as [12, 23]. In contrast with these projects, this article considers the use of discrete speech in the much more demanding Category D testing environment.

Our second result is a demonstration of the habitability of discrete speech for humans. It is quite common for speakers to adapt their vocabulary, syntax, and manner of speaking to the listener, as in speech to children, to the elderly, or to professional colleagues (see, e.g., [1, 36]), and the question is whether or not people will be willing to use discrete speech and limited vocabulary in order to communicate with machines. Although mixed results are reported in this regard in [17], which describes a series of experiments with a "simulated listening typewriter," no ultimate conclusions can be drawn for long-term interactions in an applications environment. The work reported here indicates that people are able to speak in this mode for relatively short time periods to effectively solve problems and give a reasonably positive report of their experiences. It appears that in the immediate future many voice input applications will involve discrete speech.

REFERENCES

1. Ashburn, G., and Gordon, A.M. Features of a simplified register in speech to elderly conversationalists. *Int. J. Psycholinguistics 8–3*, 23 (1981), 7–31.
2. Barnett, J., Bernstein, M.I., Grillman, R. and Kameny, I. The SDC speech understanding system. In *Trends in Speech Recognition*. W.A. Lea, Ed., Prentice Hall, Englewood Cliffs, N.J., 1980, pp. 272–293.
3. Bell, D.W., and Becker, R.W. Achieving high throughput with high accuracy in noise. In *Proceedings of the Voice Data Entry Systems Applications Conference*. AVOIS. Sept. 1982.
4. Biermann, A.W., and Ballard, B.W. Towards natural language computation. *Am. J. Comput. Linguist. 6*, 2 (1980), 71–89.
5. Biermann, A.W., Ballard, B.W., and Sigmon, A.H. An experimental study of natural language programming. *Int. J. Man-Machine Stud. 18*, (1983), 71–87.
6. Biermann, A.W., Rodman, R., Ballard, B., Betancourt, T., Bilbro, G., Deas, H., Fineman, L., Fink, P., Gilbert, K., Gregory, D., and Heidlage, F. Interactive natural language problem solving: A pragmatic approach. In *Proceedings of Conference on Applied Natural Language Processing*, Santa Monica, Calif., Feb. 1983, pp. 180–191.
7. Carroll, J.B., Statistical analysis of the corpus. In *Word Frequency Book*, Houghton Mifflin, Boston, pp. xxi–xxxix.
8. Chew, V. *Comparisons Among Treatment Means in an Analysis of Variance*. U.S. Dept. of Agriculture Publ. ARS/H/6, 1977.
9. Dahl, H. *Word Frequencies of Spoken American English*. Verbatim, Essex, Conn. 1967.
10. Damerau, F.J. The transformational question answering system operational statistics 1978. Rep. RC 7739, IBM T.J. Watson Research Center, Yorktown Heights, N.Y., 1979.
11. Dersch, W. Direct voice commands on a modern nuclear submarine. In *Proceedings of the Voice Data Entry Systems Applications Conference*. AVOIS. Sept. 1982.
12. Dixon, N.R. and Martin, T.B. (Eds.) *Automatic Speech and Speaker Recognition*. IEEE Press, New York, 1978.
13. Erman, L.D., Hayes-Roth, F., Lesser, V.R., and Reddy, D.R. The Hearsay-II speech understanding system: integrating knowledge to resolve uncertainty. *ACM Comput. Surv. 12*, (1980), 213–253.
14. Fink, P.K. The acquisition and use of dialogue expectation in speech recognition. Ph.D. dissertation, Dept. of Computer Science, Duke University, Durham, N.C., Dec., 1983.
15. Fink, P.K., Sigmon, A.H., and Biermann, A.W. Computer control via limited natural language, *IEEE Trans. Syst., Man Cybern. SMC-15*, 1985, to appear.
16. French, N.R., Carter, C.W., and Koenig, W. Jr. The words and sounds of telephone conversation. *Bell Syst. Tech. J. 9*, (1930), 290–324.
17. Gould, J.D., Conti, J., and Hovanyecz, T., Composing letters with a simulated listening typewriter. *Commun. ACM 26*, 4 (1983), 295–308.
18. Hendrix, G., Sacerdoti, D., Sagalowicz, D., and Slocum, J. Developing a natural language interface to complex data. *ACM Trans. Database Syst. 3*, 2 (1978), 105–147.

19. Howes, D. A word count of spoken English. *J. Verbal Learn. Verbal Behav. 5*, (1967), 572–606.
20. Itakura, F. Minimum prediction residual principle applied to speech recognition. *IEEE Trans. Acoust., Speech, Signal Process. ASSP-23*, (1975), 67–72.
21. Kaneko, T., and Dixon, N.R. A hierarchical decision approach to large vocabulary discrete utterance recognition. *IEEE Trans. Acoust., Speech, Signal Process. ASSP-31*, 5 (1983), 1061–1066.
22. Kucera, H., and Francis, W.N. *Computational Analysis of Present-Day American English*. Brown University Press, Providence, R.I., 1967.
23. Lea, W.A. *Trends in Speech Recognition*, Prentice-Hall, Englewood Cliffs, N.J., 1980.
24. Lea, W.A. Problems in predicting performance of speech recognizers. In *Proceedings of the Workshop on Standardization for Speech I/O Technology*. The Institue for Computer Sciences and Technology, National Bureau of Standards, Mar. 1982.
25. Lowerre, B.T. The HARPY speech recognition system. Ph.D. dissertation, Carnegie-Mellon University, Pittsburgh, Pa., 1976.
26. Mantinband, J.Y. Cyberphobia: Understanding why people fear computers. Master's thesis, Dept. of Industrial Engineering, North Carolina State University, Raleigh, 1983.
27. Martin, T.B. Practical applications of voice input to machines. In *Proceedings of IEEE 64*, (1976), pp. 487–501.
28. Miller, H.G., Hershman, R.L., and Kelly, R.T. Performance of a natural language query system in a simulated command control environment. Tech. Rep., Advanced Command and Control Architectural Testbed Facility, U.S. Navy, 1978.
29. Nye, J.M., Vocie integration: The critical mass. In *Proceedings of the Voice Data Entry Systems Applications Conference*. AVIOS, Sept. 1982.
30. Nye, J.M. Human factors analysis of speech recognition systems. *Speech Technol. 1*, 2 (Apr. 1982), 50–57.
31. Petrick, S.R. On natural language based computers. In *Linguistic Structures Processing*. A. Zampolli, Ed., North-Holland, Amsterdam, 1977.
32. Pierrel, J.M. Etude et Mise en Oeurre de Contraintes Linguistiques en Comprehension Automatique du Discours Continu. Thesis, University of Nancy, 1981.
33. Pols, L.C.W. Real time recognition of spoken words. *IEEE Trans. Comput. C-20*, (1971), 972–978.
34. Schmandt, C., and Hulteen, E.A. The intelligent voice-interactive interface. In *Proceedings of Human Factors in Computer Systems*, Gaithersburg, Md., 1982, pp. 363–366.
35. Sekey, A. Building a model for large vocabulary isolated word recognition. *Speech Technol. 1*, 2 (Sept. 1984), 71–81.
36. Snow, C.E., and Ferguson, C.A., Eds. *Talking to Children: Language Input and Acquisition*, Cambridge University Press, Cambridge, 1977.
37. Thompson, F.B., and Thompson, B.H. Practical natural language processing: the REL system as prototype. In *Advances in Computers, Vol. 13*, M. Rubinoff and M. Yovits, Eds., Academic Press, New York 1975, pp. 109–168.
38. Waibel, A. *Towards Very Large Vocabulary Word Recognition*. Carnegie-Mellon University Computer Science Department Speech Project, Nov., 1982.
39. Walker, D. *Understanding Spoken Language*. North-Holland, New York, 1978.
40. Wolf, J.J., and Woods, W.A. The HWIM speech understanding system. *IEEE International Conference Record on Acoustics, Speech, and Signal Processing*, (1977), pp. 784–787.

CR Categories and Subject Descriptors: H.1.2 [**Information Systems**]: User/Machine Systems—*human factors*; I.2.1 [**Artificial Intelligence**]: Applications and Expert Systems—*natural language interfaces*; I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*speech recognition and understanding*

General Terms: Human Factors, Measurement

Additional Key Words and Phrases: natural language processing, speech recognition and understanding, voice interactive systems, user interfaces, human-computer interaction, dialogue.

Received 4/84; revised 11/84; accepted 12/84

Authors' Present Addresses: Alan W. Biermann, Department of Computer Science, Duke University, Durham, NC 27706. Robert D. Rodman, Department of Computer Science, North Carolina State University, Raleigh, NC 27607. David C. Rubin, Department of Psychology, Duke University, Durham, NC 27706. J. Francis Heidlage, Department of Physiology, Duke University, Durham, NC 27706.