Designing the Human Machine Interface in the ATIS Domain

B. Bly P. J. Price S. Park S. Tepper E. Jackson V. Abrash

SRI International
333 Ravenswood Ave.
Menlo Park, CA 94025

Abstract

Spoken language systems for the near future will not handle all of English, but, rather, will be limited to a domain-specific sub-language. Accurate modeling of the sub-language will depend on analysis of domain-specific data. Since no spoken language systems currently have a wide range of users, and since variability across users is expected to be large, we are simulating applications in which a large population of potential users can be sampled. The data resulting from the simulations can be used for system development and for system evaluation. The application discussed here is the air travel domain using the Official Airline Guide (OAG) reformatted in a relational structure.

This study assesses the effects of changes in the simulations on the speech and language of the experimental subjects. These results are relevant to both the experimental conditions for data collection and the design of the human interface for spoken language systems. We report here on five experiments: (1) the effect of longer instructions with examples vs. shorter instructions, using our earlier data collection system, (2) a baseline experiment using a functional equivalent of the data collection effort at Texas Instruments (TI), (3) the use of a more specific version of the scenario used in the baseline experiment, (4) the use of a short, simple familiarization scenario before the main scenario, and (5) in addition to the short familiarization scenario, the use of a finite vocabulary with rejection of sentences with extra-lexical items.

Introduction

The data reported here are part of an endeavor whose goal is to design an appropriate human-machine interface by examining various parameters in a simulated interaction involving air travel planning. The design of the system is such that either a spoken language system (SLS) or a simulation of one can be inserted between the user and the relational database version of the Official Airline Guide data for North American flights and fares. In this way we can gather data for development and evaluation of both the SLS and the user interface.

Perhaps the greatest source of variability in the system is that across subjects. Individuals differ greatly in their language skills, in their problem solving skills, and in their attention spans. It is therefore important to sample a variety of subjects from the relevant population. Individuals are also very adaptable. In many cases, it may be easier to rely on subject adaptability than to try to find technological solutions. However, the dimensions along which humans might adapt are largely unknown for spoken language interfaces. Thus, the simulations provide us with a mechanism to test experimentally various interface strategies that may be appropriate for SLS technology as it develops.

We describe here five experiments aimed at answering various questions about the interface. Our first experiment, the only one reported here that was not based on a functional equivalent of the TI data collection system, investigated the effect of a long set of instructions with examples compared to a shorter set with no examples. The goal of this study was to investigate how much one "poisons the data" by using such examples. The next four experiments were based on either a functional equivalent of the TI system, or a minor variation:

- To serve as a baseline experiment to compare our results to those of TI, and to serve as a control for the other experiments, we collected data in a fashion that imitated the TI system as much as possible.
- To investigate the effects on yield that might result when subjects interpret what a vague scenario might mean, we modified the scenario to fill in details that were unspecified in the original.
- To investigate the first session effect, which was large in our earlier work, we used a simple, short (about 5-minute) familiarization scenario.
- To investigate how well subjects might adapt to a fixed vocabulary, we used a short familiarization scenario, gave subjects a list of about 1000 words, and gave error messages for utterances with words not on that list.

Data Collection Conditions

Except for the first experiment, which was carried out before the functional equivalent of the TI data collection

system had been completed, our aim was to imitate as well as we could the system used by TI for data collection. In particular, we have used the same data from OAG formatted in the same relational structure; the same tool for the "wizard" (NLParse) and accompanying NLParse grammar; the same relational database (Oracle) and interface to NLParse; the same set of tools for communication among subject, wizard, and transcriber; the same subject and experimenter instructions; and the same formatting of tables and other objects displayed on the screens (controlled by Oracle). We used only one of TI's scenarios, planning a family reunion involving family members of various types.

Our data collection differed from that of TI in a few ways that we felt were either unavoidable or unimportant for the resulting data. We are aware of the following differences: our A/D system uses a NEXT machine; our push-to-talk mechanism writes out a time stamp for push and for release (this allows us to calculate the time spent speaking, waiting for an answer and thinking before making the next query, which the TI system does not allow); instead of the color coding used by TI, we use a "ready" prompt when the system is ready to accept speech, a "listening" prompt when the subject is pushing the mouse button, and a "processing" prompt after the subject releases the button and before the answer is sent. We offered a free "DECIPHER" T-shirt to participants in an experimental session.

Data Analysis

Each session was timed from beginning to end, the training scenarios were timed, and the delay until the subject initiated the first utterance was timed. The numbers of words and utterances produced per session were counted, as were the numbers of words and utterances produced during the training scenario. A time stamp was automatically recorded each time the subject used the push-to-talk button, each time a transcription was sent, and each time a response was sent to the subject's screen. This allowed us to determine the average time the subject took after receiving an answer and before formulating a query (thinking time), the average time the subject held down the push-to-talk button (speaking time), and the average time it took the wizard and the wizard's assistant to send the transcription and database response to the subject's screen (subject waiting time). The average number of words per utterance, the average vocabulary size per subject, and the number of sentences outside the restricted vocabulary used in the Fixed Vocabulary Condition were counted. We also counted the number of cancellations subjects used per session, and the number of error messages sent. After the session, all subjects filled out an eleven-item questionnaire designed to assess their subjective impressions of the system and their satisfaction with their interaction with the system. Analyses of these measures were completed for the ten subjects in each of the four conditions that were based on the TI data collection system.

For the word counts, we used the .nli files (see [2]), and used functions to reformat the data so that, for example "845" would count as three words rather than one. Other, similar changes were made to regularize the spellings.

Condition 0: Long Instructions

This condition is the only one that is not based on the TI data collection system; it is based on the system described in [1]. We describe it briefly here since the results were part of the motivation for the two training conditions described below.

This experiment tested the effect of subject instructions on the language produced by the subjects. Two sets of instructions were used: one that included ten grammatical and parsable utterances as examples, and one that included no examples. In all other respects they were identical. Based on previous work, we expected a large effect of experience with the system, so subjects were asked to perform two tasks, and performance was compared across the two tasks as well as between the two sets of instructions. 208z We found a strong interaction between the type of instructions given and the amount of experience the subject had with the system; that is, on a subject's first task, those who received long instructions behaved like the more experienced, second-task subjects on the measures used in the previous study. They also used more complete sentences and did not show the pattern of short, choppy, telegraphic speech demonstrated by the subjects who received a short set of instructions. It is possible, then, to affect the speech the subject addresses to an SLS by providing examples. It is important to note that the effects of longer instructions and additional experience with the system were not additive: new users appear to need either detailed instructions or additional practice time but not both.

The data collected in this experiment was different in important ways from data collected and reported by TI. The sentences, especially those produced by subjects not given examples, were shorter (an average number of 7.4 words per utterance compared to about 12 for the TI data). However, due to the many differences between this interface and that used by TI, it was impossible to reliably attribute these differences to any specific causes. We therefore designed a series of minor modifications of the TI version, as described below.

Condition 1: TI Equivalent

The goal of the "TI" Condition was to establish that our data collection system was a functional equivalent of the TI system, and then to serve as a baseline for the subsequent conditions. We tried to conform as closely as possible to TI's methods, physical setup and materials. In this condition, subjects were read a set of instructions identical to the instructions used by TI, the task they were asked to perform was one of the TI scenarios, and

	TI	SRI-TI
No. utterances	26.2	23.5
No. words	305	298
Words/utterance	11.6	12.7
No. unique words/subj.	83	81
No. unique words/cond.	286	296
Time between utterances	90 sec.	89 sec.

Table 1: SRI-TI Condition Compared with TI Data

the wizard was familiar with NLParse and had practiced, using the transcription and query data released by TI.

The data from our TI Condition seems to match TI's released data very well. As shown in Table 1, the various measures made are all very similar.

Perhaps the most striking difference between TI's data and SRI's in the TI Condition appeared in an analysis of word frequency. We were astonished that the frequencies were so different for "show" (75 occurrences in TI's data vs. 8 in ours). Similar discrepancies showed up for the words "me", "nonstop" and "flights". We then realized that the sentence used by TI as an example demonstrating the use of the mouse and the formatting of the tables, "Show me all the nonstop flights from Atlanta to Philadelphia", had a profound effect on the resulting data (though, of course, these utterances from each speaker were not used in the analysis). In our data collection, we asked the subject to read the first sentence of the scenario while we verified the recording procedure and demonstrated the push-to-talk button.

Condition 2: Task Specificity

We found, in examining both data released by TI and our own data in the TI Condition, that it was often hard to tell how a subject had interpreted a given task, and even which task was being performed. The data could be more valuable if we could ascertain whether and how well the subject completed the task. We also thought that subjects would be more cooperative and the task would be more realistic if they were concentrating on solving the task rather than on exploring the limits of the system. In addition, we suspected that some time might be wasted while the subject tries to figure out what the task is.

To eliminate the effect of individual interpretation of the task and to standardize the task across all subjects, we ran a "Specific Task" Condition. In this condition, subjects were given the same instructions as in our TI Condition. The task they were asked to perform, however, while structurally the same as the tasks performed by TI's subjects and by our own subjects in the TI Condition, was more specific. Instead of leaving the interpretation of certain aspects of the task to the subjects (for instance, find a flight for a person with an "adventurous" lifestyle), we set explicit constraints (find an airplane that holds the fewest number of passengers). In addition, instead of choosing any cities from the database to complete the task, subjects were assigned the origin and destination cities. Each of the ten subjects in this condition used a different set of four cities, determined randomly from the set of cities in the database. In all other aspects, this condition was identical to the previous condition.

We found no significant differences on any of our measures between the subjects in our TI Condition and our Task Specificity Condition. It may be that any benefits gained by subjects not being required to fill in the details themselves were offset by the fact that assigning random cities does not work as well as when subjects pick the cities themselves. For example, several of our subjects had difficulties because they did not realize that Dallas and Fort Worth shared an airport. Subjectively, however, it did appear that subjects completed the assigned task, whereas in the TI Condition, many subjects gave up or quit before fulfilling the various parts of the task required by the scenario. We are working to develop objective measures of this subjective impression of the "dialogue" quality of the collected utterances.

Condition 3: Familiarization

Our past data collection efforts showed a large effect of user experience in human-human interactions and in experimental human-machine interactions [1]. In both conditions, the more domain-experienced speakers produced fewer words, fewer false starts and fewer filler words than did the less-experienced speakers. In addition, subjects elicited fewer error messages in their second scenarios compared to their first. Further, the dramatic effect of one sentence read by all subjects at TI shows just how adaptable subjects can be, at least in an initial session.

In the "Familiarization Condition", after reading the same instructions as in the other conditions, the experimenter stayed in the room with the subject and answered any questions the subject had in finding a single one-way flight between San Francisco and Dallas. The experimenter responded to questions including those regarding the kind of requests the system could handle, the kind of information in the database, and the push-to-talk button. The experimenter also provided possible explanations for any error messages the subject received during the training scenario. The familiarization scenario remained constant across all subjects, although the scenarios that constituted the main task varied among subjects as described in the Task Specificity Condition above. The average length of a training scenario was 6.57 minutes.

Among the various conditions we ran, the largest effect by far was that of the familiarization scenario. As shown in Table 2, subjects who used familiarization sce-

	No Familiarization	With Familiarization
Task time	40 min.	23 min.
Utterances/Task	24	17
Words/Task	276	146
Words/Utterance	12.2	8.7
Format queries	25%	13%
Errors	3.9	1.2
Cancellations	3.8	1.6
Thinking time	46 sec.	34 sec.
Speaking time	8.2 sec.	6.9 sec.
Waiting time	42 sec.	39 sec.

Table 2: Comparison of Conditions with and without Familiarization Scenario

	1	2	3	4
Unique words/subject	81	89	83	67
Unique words/condition	296	344	270	219
Extra-lexical items, No. words	66	80	61	0
Extra-lexical items, No. sentences	74	205	138	0
(percent sentences)	(31)	(81)	(87)	(0)
Vocabulary errors	0	0	0	3.8
Other errors	3.7	4.2	1.8	0.6
Task Time (min)	37	43	22	24

Table 3: Comparison of Condition 1 (SRI-TI), 2 (Task Specificity), 3 (Familiarization Scenario), and 4 (Finite Vocabulary).

narios took significantly less time to complete the main task (23.2 vs. 39.9 minutes, $p < .01$) and used significantly fewer words to complete the task (276 vs. 146, $p < .01$) than subjects in the other two conditions. The difference between the number of utterances produced by the two groups was not significant, however (24.4 vs. 17.2, $p > .05$), while the number of words per utterance used by subjects in the training conditions was fewer (8.7 vs. 12.2, $p < .01$). Subjects in the familiarization conditions also received fewer error messages per utterance produced (.07 vs. .13) and asked fewer questions concerning the meanings of table headings (13% of all queries, compared to 25% for subjects with no familiarization scenario).

Condition 4: Finite Vocabulary

Earlier work concerning the vocabulary used by subjects and the percent of new words introduced in each session suggested that expert human-machine users could potentially adapt to a restricted vocabulary and still maintain efficiency [1]. In order to test whether subjects would adapt to a restricted vocabulary, we slightly modified our system to accept only a limited vocabulary from the subjects. The wizard's assistant, instead of being provided with a normal spell-checker, used a spell-checker that contained only a subset of approximately 1000 most frequently used words, based on the data released by TI in distributions 1-4 (prepilot data plus NIST Release 1). Subjects were made aware of this restriction in the instructions and were provided with a list of acceptable words. If they used a word outside the vocabulary, they were sent the message: "You have used a word outside the system's vocabulary. Try rephrasing your request." In all other respects, this "Fixed Vocabulary" Condition was identical to the Familiarization Condition (i.e., subjects in this condition were given a familiarization scenario and performed a constrained task).

If we compare the subjects who received a familiarization scenario but were unlimited in vocabulary and those who received a familiarization scenario but were limited to a 1000-word vocabulary, we find that the error messages received by the latter group for using out-of-vocabulary items is higher. During the familiarization session, they received an average of 2.0 error messages of this kind, and an average of 3.8 messages of this kind for the main task. When added to the other error messages they received, this gave them a slightly higher number of total error messages received than subjects in the comparable but unlimited-vocabulary condition (4.4 vs. 1.8). The mean number of error messages received by the group was not, however, different from the mean number of error messages received by subjects in either of the non-familiarization scenario conditions. In addition, there is evidence for the adaptation of subjects to a fixed vocabulary as indicated in Table 3. This table indicates that with a short familiarization session and consistent feedback one can dramatically affect the number of unique words used by the subject, the number outside a fixed set, and the number of sentences with such "extra-lexical" items, without increasing the total time to complete the task. The discrepancies between the number of "extra-lexical" items and the number of

sentences in which they occur arise because some subjects will use a given lexical item in many subsequent sentences once it has "worked".

Discussion

In addition to replicating the results released by TI, using a setup similar to TI's, we have shown the effect of altering various aspects of the experimental setup, including scenario specificity, subject familiarization and restricting the vocabulary.

We believe that our results indicate that we have succeeded in implementing a functional equivalent of the TI data collection system. The one major exception to this claim is the observed discrepancy in the word frequency distributions. This discrepancy can be remedied by avoiding any sample sentences from the domain while instructing subjects.

In assessing scenario specificity, we found no differences on either yield measures (time to complete task, utterances per task, words per task, etc.) or on quality measures (error message rates, cancellation rates) between subjects in the unconstrained task condition and those in the constrained (specific) task condition. In light of this, one might argue for adopting specific scenarios on the basis of the benefits gained by knowing subjects are interpreting the task the same way (in effect, are performing the same task) and by obtaining data useful for both analysis of isolated queries and of dialogue.

Our most significant results pertain to subject familiarization. In two separate experiments using two very different interfaces and procedures, we demonstrated the impact of subject familiarization with the system: subjects less familiar with the system produced longer utterances, needed more time to complete the task, and produced fewer utterances per subject hour. The time to familiarize subjects with the system (5 to 6 minutes) was short relative to the gains in subject efficiency (17 minutes saved on average in subject time to complete task).

Our Fixed Vocabulary Condition showed that subjects can adapt quickly to a restricted vocabulary without increasing task time: subjects in the Fixed Vocabulary Condition did not take longer to complete the task or to plan each utterance than those in the unlimited-vocabulary conditions, so the constraint doesn't appear to slow them down unnaturally or lower the yield of the experimental session. It is worth noting that these subjects showed significant improvement in the number of out-of-vocabulary error messages received during the main task (3.8 in 24.29 minutes) as compared to the training scenario (2.0 errors in 7.27 minutes). This supports the position that subjects can adapt to using a limited vocabulary. This result may be very important in the development of scalable technologies that will fit on a variety of platforms.

We found no systematic differences in the answers subjects provided to the questionnaire we presented to them after the session. The subjective experience of the subjects in the various conditions, then, seems to have been about the same.

The goals of designing an appropriate spoken language system can sometimes conflict with the goal of collecting data for evaluation of spoken database queries. That is, some major causes of errors (e.g., out-of-vocabulary items, out-of-domain queries) may disappear with a small amount of either detailed instruction or subject familiarization. However, we are convinced that it is possible to find ways of coordinating the two endeavors. For example, the needs of both dialogue analysis and of query-answer pairs for evaluation can be met using a more specific scenario; the needs of restricted vocabulary can be met by providing consistent feedback; and the large effect of subject familiarization can be addressed by spending a short time in the room with the subject to answer questions as the subject works on a task.

We plan to continue these experiments to help us design an appropriate human-machine interface. In our next set of experiments we will include a revised grammar for NLParse that reduces the number of words the wizard needs to produce by about 35% (on "cheapest" constructions it can reduce the number of words to about a quarter of the number that would be needed without the modification). Other experiments we are planning include the reformatting of tables sent by Oracle (the high percentage of queries concerning the meanings of various column headings indicate that much could be done to improve the user interface in this area), and some variations on the use of push-to-talk mechanism. We will also be running repeat subjects to test the effect of longer use of the system on the resulting data.

Acknowledgements

The authors gratefully acknowledge TI and Charles Hemphill, in particular, for helping us to get a license to NLParse, for providing much of the related software and data, and for helpful responses to our endless pleas for assistance. We also thank CMU for providing recording and playback software for the NEXT machine. This research was funded by DARPA under Office of Naval Research contract N00014-90-C-0085.

References

[1] J. Kowtko, P. Price and S. Tepper, "Data Collection and Analysis in the Air Travel Planning Domain," DARPA Speech and Natural Language Workshop, October 1989.

[2] C. Hemphill, J. Godfrey,, and G. Doddington, "The ATIS Spoken Language Systems Pilot Corpus," this volume.