Spontaneous Speech Collection for the ATIS Domain with an Aural User Feedback Paradigm

Christine Pao and Jay Wilpon

AT&T Bell Laboratories
600 Mountain Ave. Office: 2D-464
Murray Hill, NJ 07974

ABSTRACT

This paper describes the AT&T ATIS data collection system, with emphasis on the development of the speech-in, speech-out interaction paradigm. The ATIS task involves providing air travel information to a user in the context of a interactive dialogue. Under the AT&T interaction paradigm, information retrieved from a travel information database in tabular form is automatically transformed into sentences, which are spoken to a user by a speech synthesizer. To date, we have collected over 1800 sentences from subjects who used the system to solve travel planning scenarios. We present a comparison of the ATIS data collected at AT&T with the ATIS data collected at other sites (BBN, CMU, MIT, and SRI), and discuss what we have learned in this preliminary effort.

1. INTRODUCTION

In support of our research toward developing telephone-based spoken language systems, we have joined BBN, CMU, MIT, and SRI in collecting speech and language data for the ATIS (Air Travel Information Service) domain. The task of a spoken language system can be broken into three parts. First, the system must understand and interpret what the user says in the context of the human-machine dialogue. If the user's request is understood, the system must retrieve the requested information. The information source for the ATIS domain is a relational database that represents a 10-city subset of the OAG, or Official Airline Guide. Finally, the system must convey the retrieved information to the user in an appropriate format, or provide some other cooperative response, such as a request for more information or an error message. In the AT&T ATIS data collection system, the first and second portions of the task are handled by components of the MIT ATIS system [1]. Our efforts have been directed toward dealing with the third part of the task: information presentation and system feedback under a speech-in, speech-out interaction paradigm. Our goal is to begin to address the problems of dialogue and information control which will affect the performance of interactive spoken language systems.

In this paper, we will first describe our data collection system and data collection procedure. Then, we will present analyses of AT&T ATIS data and of ATIS data collected at other sites. Finally, we will discuss what we have learned in this preliminary effort.

2. DATA COLLECTION

As at most of the other sites collecting ATIS data, data were collected from subjects' interactions with a partially simulated spoken language system. As in the MIT data collection setup, a human experimenter is substituted for the speech recognition component of the system to provide a transcription of the subject's speech for the natural language (NL) component. In this section, we discuss the system's development, describe the system hardware, and describe the collection procedure.

2.1. System Development

The AT&T data collection system was designed to closely simulate a real, telephone-based, human-machine interaction. Building on the framework of the MIT ATIS collection system, we directed our development effort towards controlling the presentation of information retrieved from the database, providing feedback to the user on the state of the system and the discourse, and exploring areas where system initiative would help users achieve their goals efficiently. We also modified the system control loop to transfer recording control away from the subject.

Information Presentation. The collection systems at all other sites make use of a visual display to present information retrieved from the database in a tabular format. Because of our choice of an audio interaction paradigm, the AT&T system does not present information in a tabular format, but instead translates the retrieved information into sentences. In some cases, information is converted into sentences using an entry-to-phrase, template-based approach (Figure 1). When more information is retrieved from the database than can be reasonably presented by the template-based approach, summarization functions are used to present some subset of the information (Figure 2). In other cases, specialized functions are used to avoid excessive repetition (Figure 3) or to select information based on the discourse history.

```
Sentence:  I'LL TAKE DELTA FLIGHT NINE SEVENTY FIVE
  Entry Type         Entry    Sentence Fragment
  FLIGHT_ID          105595
  AIRLINE_CODE       DL       delta
  AIRLINE_FLIGHT     975      flight nine seventy five
  FROM_AIRPORT       BOS      from Boston
  TO_AIRPORT         ATL      to Atlanta
  DEPARTURE_TIME     1520     departs at three twenty P M
  ARRIVAL_TIME       1804     and arrives at six oh four P M
  STOPS              0        with no stops
ATIS: delta flight nine seventy five from Boston to
Atlanta departs at three twenty P M and arrives at six oh
four P M with no stops.
```

Figure 1: Template-based conversion of flight information to sentence format.

```
Sentence:  I WANT TO GO FROM BOSTON TO ATLANTA ON MONDAY
Table:
  FLIGHT_ID    AIRLINE_CODE    ...    DEPARTURE_TIME    ...
  105584           DL          ...         630          ...
  105586           EA          ...         700          ...
  105588           DL          ...         815          ...
     :
Summary:  There are flights departing between six thirty
A M and eight twenty four P M.
ATIS: There are seventeen flights from boston to atlanta
on Monday August nineteenth.
There are flights departing between six thirty A M and
eight twenty four P M.
What time would you like to go?
```

Figure 2: Output of the table summarizer. The summary is one part of a three sentence response to the user.

```
Sentence:  WHAT AIRLINES FLY FROM BOSTON TO ATLANTA
Table:
         AIRLINE NAME            AIRLINE CODE    FROM    TO
  DELTA AIR LINES, INC.               DL         BOS     ATL
  EASTERN AIR LINES, INC.             EA         BOS     ATL
  USAIR                               US         BOS     ATL
ATIS: The airlines with service between Boston and
Atlanta are Delta, Eastern, and U S Air.
```

Figure 3: In this example, a table with three rows is compressed into a single sentence.

The information presentation component of the system was developed with two goals in mind. The first was to present information so it could be easily understood. Toward this end, the presentation component was developed to format information into coherent sentences, to expand or hide all codes and abbreviations, and to maximize the intelligibility of the speech synthesizer output. The second goal was to minimize the amount of irrelevant information presented to the user. Towards this goal, the presentation component includes the above mentioned facilities for summarizing, compressing and filtering information retrieved from the database.

```
Sentence:  I WANT TO GO ABOUT 3 P M
ATIS: There are no flights leaving between two forty five
P M and three fifteen P M.
The next earliest flight is eastern flight six forty five
departing at two twenty one P M.
The next latest flight is delta flight nine seventy five
departing at three twenty P M.
Please refer to these flights by flight number or
departure time.
```

Figure 4: AT&T system initiative.

System Feedback. The MIT ATIS system provides feedback to the user on the state of the discourse in the form of text and synthesized speech. However, the supporting text produced by the MIT system is intended to complement and to direct the user's attention to a tabular display. This capability was modified to complement the summarization facility mentioned above. In the example shown in Figure 2, the MIT system would generate the text "Here are the flights from Boston to Atlanta" to accompany the table listing seventeen flights, while the AT&T system would generate "There are seventeen flights from Boston to Atlanta" to accompany the following summary sentence. The system error responses (NL failure, database access failure, etc.) were also modified to fit the audio feedback paradigm.

System Initiative. The MIT system takes initiative in two contexts: when the system does not have enough information to access the database, and when guiding a user through the flight booking process [3]. The AT&T system takes initiative in an additional context, when the subject is selecting a flight on the basis of departure or arrival time. First, the system prompts the user for a departure time if the departure time is summarized, as in Figure 2. Second, the system volunteers the next earliest and next latest flight when the subject requests a flight at a certain time, and there isn't one. An example is shown in Figure 4. This second capability was developed to address a problem that was causing a great deal of user frustration. Because the system would not provide complete flight information for more than three flights, subjects were forced to play a guessing game to find out flight departure times. The need for this type of system initiative is a result of the limits imposed by the interaction paradigm. The more a system restricts the flow of information, the more assistance it must provide to help the user access the information.

Recording Control. At all the other ATIS data collection sites, the subject controls the recording process using a push-to-talk or push-and-hold to talk mechanism. We chose not to use such a subject-controlled recording mechanism in order to more closely simulate an actual telephone dialogue. Instead, the experimenter who transcribed the subject's speech also controlled the start and end of recording from the keyboard. The control loop was designed to keep the interaction flowing as smoothly and efficiently as possible in the hope of eliciting more natural speech from our subjects. Many subjects were initially uncertain about when to start and stop talking, but most of them adjusted to the interaction after the first scenario. Some effects of experimenter-controlled recording on subjects' speech are discussed in section 3.4.

2.2. Recording Environment and System Hardware

Data were collected in a walled-off corner of a computer laboratory. The subjects were seated at a desk with a telephone, and provided with paper and writing implements. All system feedback to the subject was provided over the telephone by the AT&T TTS speech synthesizer. Speech data were captured simultaneously using (1) a Sennheiser HMD-410 close-talking microphone amplified by a Shure FP11 microphone-to-line amplifier, and (2) a standard carbon button microphone (in the telephone handset) over local telephone lines. Digitization was performed by an Ariel Pro-Port A/D system.

2.3. Data Collection Procedure

Before a recording session began, the experimenter provided the subject with a brief verbal explanation of the task and a page of written instructions. The subject also received a summary of the task domain and two sets of travel planning scenarios. The first set of scenarios included a number of simple tasks (referred to below as "short scenarios") and the ATIS common scenario (used at all five ATIS data collection sites). The second set contained more complicated tasks (referred to below as "long scenarios"), and subjects were permitted to attempt to book flights while working on these scenarios. Initially, the subjects selected which scenarios they wanted to try; because of problems with uneven scenario distribution, the experimenter began selecting an initial set of scenarios (two short, one long) for each subject. Subjects were asked to speak as they would to a human being, and to speak in single sentences. They were not told that someone was listening to them and typing in what they said until after the entire recording session was over. A complete session lasted about an hour, including initial instruction, a two part recording session with a five minute break, and a debriefing questionnaire.

During the recording session, the experimenter listened to the subject's speech and the system's response. The system initiated the dialogue with the prompt, "I'm ready to begin a scenario," and responded after every utterance with information or an error message. An example of a typical series of interactions is given in Figure 5. The experimenter controlled recording from the keyboard, starting recording as soon as the system response ended, and stopping recording when the subject appeared to have completed a sentence. The experimenter was asked to transcribe exactly what the subject said, excluding false starts. However, because of (perceived) pressure on the experimenters to get answers to the subjects, especially after repeated system failure, the session transcriptions sent to the interaction log files were not always accurate. Most of the time, the experimenter interacted with the subject only through the system. However in cases of complete system failure and severe subject confusion, the experimenter could communicate directly with the subject, either by sending a message through the speech synthesizer, or by speaking directly to the subject.

Subjects for data collection were recruited from local civic organizations, and collection took place during working hours. As a result, 82% of the subjects were female, and subjects ranged in age from 29 to 77, with a median age of 55. Approximately 60% of the subjects came from the New York City dialect region; all were native speakers of English. In return for each subject's participation, a donation was made to the civic organization through which he or she was recruited.

Four summer students served as experimenters for almost all of the data collection sessions. They were trained for two weeks during pre-collection system debugging. The system was debugged and intermittently upgraded during and after the 2 1/2 week collection. All of the data was then transcribed and submitted to NIST for distribution.

3. COMPARATIVE ANALYSES

In the following section, we present analyses of data collected at AT&T and the four other data collection sites. We will pay particular attention to comparing the AT&T data with the MIT data, since the two collection systems had identical natural language and database access components, and the two sites used the same travel planning scenarios.

3.1. General Characteristics

Table 1 presents some statistics which can be used to characterize each site's collection system and procedure. The statistics were computed for each speaker, and the speaker values averaged for each site. AT&T is in the middle of the pack in utterances per scenario, collection rate (number of utterances collected divided by recording session duration), and human-machine interaction time

```
Scenario: "Determine the type of aircraft used on a
flight from one city to another that leaves before (or
after) a certain time of the day."

ATIS: I'm ready to begin a scenario.
Subject: BEGIN SCENARIO THREE
ATIS: Okay, go ahead.
Subject: <truncated utterance>
ATIS: Could you repeat that please? I didn't catch
      everything you said.
Subject: YES I WOULD LIKE TO FIND WHAT FLIGHTS ARE
         AVAILABLE FROM ATLANTA TO BOSTON ON THE
         MORNING OF JULY SEVENTH
ATIS: I'm confused. Could you try to rephrase your
      question?
Subject: I WANT TO FLY FROM BOSTON TO ATLANTA
ATIS: There are seventeen flights from boston to
      atlanta.
      There are flights departing between six thirty A M
      and eight twenty four P M.
      What time would you like to go?
Subject: SIX THIRTY A M
ATIS: There is one flight from boston to atlanta leaving
      six thirty A M.
      delta flight two twenty five from Boston to Atlanta
      departs at six thirty A M and arrives at nine
      thirteen A M with no stops.
Subject: WHAT TYPE OF AIRCRAFT WILL BE USED
ATIS: There is aircraft for one flight.
      delta flight two twenty five from Boston to Atlanta
      departs at six thirty A M and arrives at nine
      thirteen A M with no stops. The plane for this
      flight is a Mcdonnell Douglas M D eighty.
Subject: I WOULD LIKE TO BOOK THAT FLIGHT
ATIS: There is one flight.
      delta flight two twenty five from Boston to Atlanta
      departs at six thirty A M and arrives at nine
      thirteen A M with no stops.
      I'm sorry, I can help you with travel information,
      but I can't book flights.
Subject: END SCENARIO
```

Figure 5: A typical interaction between a subject and the AT&T data collection system.

(interval from the beginning of one utterance to the beginning of the next). The average AT&T speaker used significantly more words per utterance than the average speaker at any other site. This may be due to our subjects' response to NL system failure, which is discussed in section 3.2. AT&T subjects also had a lower average speaking rate than speakers from other sites. This may be related to the higher disfluency rate (discussed in section 3.4) and increases in the frequency of occurrence and durations of silent pauses.

3.2. NL System Failure

One of the effects of the audio interaction paridigm was a higher NL system failure rate (MIT 33.4%, AT&T 42.9%), where NL system failure is defined as the failure to completely process an utterance because it contains unknown words or fails to parse. The change in the interaction paradigm changed the NL task; since the NL system was designed based on a visual display, the NL system failure rate was expected to increase. The response of subjects to NL system failure was also affected by the change to the audio interaction paradigm. Table 2 shows the subjects' responses to NL failure at AT&T and MIT. Subjects at both AT&T and MIT spoke longer sentences and slowed their speaking rates. However, the effects of NL failure on subjects' speech are more dramatic in the AT&T data: the number of words per utterance increased by over 50% (MIT 20%), the speaking rate dropped by 15% (MIT 5%), and the utterance duration increased by over 75% (MIT 25%) when compared with utterances which did not follow an NL system error.

Variables	AT&T	MIT	BBN	CMU	SRI
# Speakers	50	136	45	27	39
# Scenarios	177	576	210	122	63
# Utterances	1886	4287	1616	1543	1055
Avg utterances/scenario	11.5	7.9	8.1	13.7	19.2
Avg utterances/hour	54.5	70.5	29.3	77.4	60.7
Avg interaction time	51.8	51.6	115.1	41.8	61.3
Avg words/utterance	12.4	9.5	10.9	9.9	8.9
Avg words/minute	118.4	175.1	158.8	139.0	139.5

Table 1: Summary of general characteristics of data from each site.

Variables	AT&T	MIT
Average words/minute:		
Post NL failure	106.8	153.3
Non-post NL failure	123.8	160.8
Average words/utterance:		
Post NL failure	14.8	10.2
Non-post NL failure	9.8	8.6
Average seconds/utterance:		
Post NL failure	8.3	4.0
Non-post NL failure	4.7	3.2

Table 2: These statistics reflect the subjects' response to system failure.

The large increase in utterance length after NL failure, combined with the high NL failure rate, is the main reason the average AT&T sentence is so much longer than the average MIT sentence, both in number of words and in duration. However, the reason behind the post-NL-failure increase in sentence length is not entirely clear. A qualitative examination seems to indicate that the system is not effectively communicating the reason for its failure. The NL system usually fails as a result of an unfamiliar, unusual, or ungrammatical syntactic construction. During the initial task familiarization, the subjects were told that the system failure was triggered by problems with a sentence's grammatical construction, and not by any type of recognition problem. Subjects were also informed of the system's discourse capabilities. Yet when a sentence failed to parse and the subject was asked to rephrase his or her request, he or she frequently responded by simply tacking on a summary of the previous discourse without modifying the syntactic structure of the original sentence. In these cases, the subjects appeared to respond to the NL failure as a discourse failure instead of a syntactic failure. Subjects did appear to adjust their speech to the constraints imposed by the NL system, as the NL system failure rate decreased from 51% in the first scenario to 39% in subsequent scenarios.

3.3. Vocabulary Comparisons

Table 3 contains statistics on the increase in lexicon size as a function of the number of sentences collected. The breakpoint of 600 sentences collected was chosen because it was the point at which the vocabulary growth rate remained less than 30 words/100 sentences for all sites. MIT has reached a terminal vocabulary growth rate of 8.7 new words/100 sentences collected; the vocabulary growth rates at the other four sites continue to decrease as the number of collected sentences increases. Figure 6 is a graph of lexicon size vs. number of sentences collected for each site.

Figure 7 shows the overlap of the different sites' lexicons. Excluding the AT&T data, the percentage of words in lexicon X that are found in lexicon Y appears to be proportional to the size of lexicon X. However, the percentages of words in the AT&T lexicon that appear in the BBN, CMU, and SRI lexicons are lower (by about 5 percentage points) than predicted, though the overlap with the MIT lexicon matches the prediction fairly well. One explanation is that, although the change in the interaction paradigm does affect the vocabulary, the effect of the change in paradigm is similar to the effects of other inter-site system variations. The AT&T system differs from the MIT system only in the interaction paradigm, but differs from the BBN, CMU, and SRI systems in other ways, in addition to the different interaction paradigm.

Variables	AT&T	MIT	BBN	CMU	SRI
Total utterances	1885	4247	1616	1543	1055
Above 600 sentences*	13.75	12.64	19.70	14.78	15.75

*For MIT, between 600 and 2000 sentences.
MIT above 1000: 8.72 new words/100 sens

Table 3: Vocabulary growth: New words per 100 sentences collected.

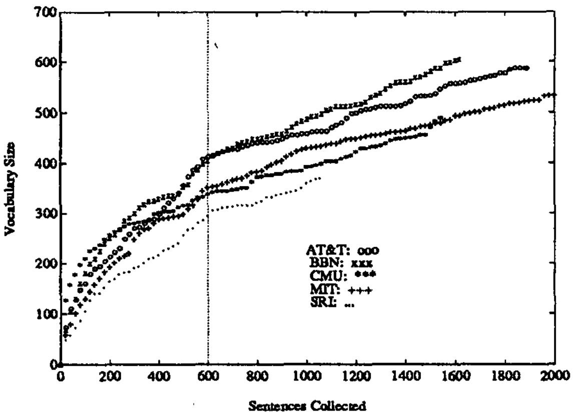

Figure 6: Number of sentences vs. vocabulary size.

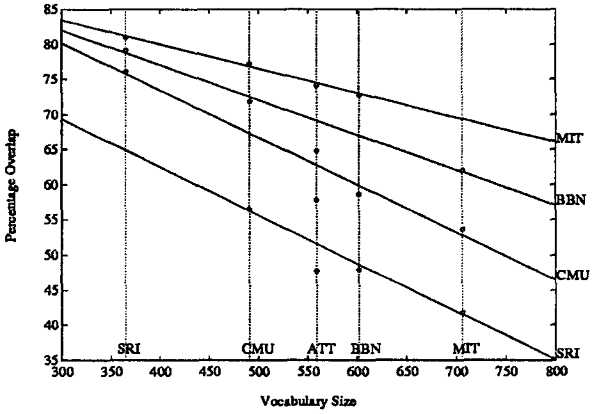

Figure 7: Percentage of inter-site vocabulary overlap.

3.4. Disfluencies

Table 4 contains statistics on the occurence of disfluencies as transcribed by each site. Partial words and word fragments are counted as lexical false starts, and verbally deleted complete words are counted as linguistic false starts, as in [2]. The high percentage of utterances containing linguistic false starts and filled pauses in the AT&T data reflect the subjects' response to experimenter-controlled recording and their uncertainty about how to interact with the system. Since the subjects did not take any initiative in starting and stopping recording, they were less likely to compose their thoughts before they began speaking. The rate of filled pauses and false starts did decrease somewhat as the subject became more comfortable talking to the system: comparing disfluency rates for first scenario and last scenario utterances, the percentage of utterances containing linguistic false starts decreased from 13% to 11%, the percentage containing lexical false starts from 8% to 7%, and the percentage containing filled pauses from 15% to 12%.

NL system failure strongly affected the rate of linguistic false starts. The percentage of utterances containing linguistic false starts increased from 9.4% after a successfully processed utterance to 14.4% after an NL system error. The absence of similar increases in the rates of lexical false starts and filled pauses indicates that the subjects' speech was disrupted primarily at the syntactic level.

% of sentences with	AT&T	MIT	CMU	BBN	SRI
linguistic false starts	11.4	3.9	1.2	2.4	4.5
lexical false starts	7.6	2.8	9.3	2.2	2.8
filled pauses	13.7	3.1	3.0	1.9	1.5

Table 4: Percentage of sentences containing various disfluencies.

4. DISCUSSION

Our initial effort in collecting ATIS domain data under a speech-in, speech-out interaction paradigm has produced some interesting results. A number of issues have come up during system development, data collection, and data analysis which need to be considered in the development of telephone-based spoken language systems and spoken language systems in general.

The system with which we collected data was not ideal. Many of the subjects were able to get around the system's limitations, but others had a great deal of trouble. As a result, some of the speech we collected sounds perfectly normal, and some sounds exceptionally unnatural and unusual, and not like normal conversational speech at all. Because of the system's inefficiency, some people found it difficult to keep track of the dialogue. Other subjects had problems because of the high system failure rate (NL and otherwise), ineffective communication of the discourse history, and confusion about system limitations. Although most subjects said they had no problems understanding the output of the speech synthesizer, it appeared that many subjects had trouble absorbing and remembering the information presented. Many potentially useful system components were only partially developed, and sometimes caused new problems. Because the bounds of the task domain as defined by the database did not match the bounds inferred by the users based on the travel planning scenarios, subjects frequently wandered out of the domain.

Developing an interactive system like ATIS under an audio-only paradigm is more difficult than developing a similar system under a less restrictive feedback paradigm. The audio interaction paradigm demands more effort on the system's part in compressing and filtering information before it is presented to the user, and in making sure all the information the user needs is easily accessible. It is more difficult to communicate system limitations and system status to the user, since information cannot be provided continuously or from more than one source at a time, and the quantity of information is limited to what a user can be expected to absorb and remember. Because the perceived waiting time is longer with no visual distractions, a system operating under an audio feedback paradigm must be efficient, so that the flow of the dialogue is maintained and the user remains attentive. There must be logical closure in the system's capabilities, and the limits of the task domain must be obvious to the user. These issues are critical in developing telephone-based systems, and many are important in the development of any interactive system.

We intend to continue the development of the ATIS system, particularly in the areas of dialogue and information control. We will also continue to collect data to support our research in telephone-based spoken language systems, and in support of the speech and language research community in general.

5. ACKNOWLEDGEMENTS

We would like to acknowledge the members of the MIT Laboratory for Computer Science Spoken Language Systems Group who provided us with the MIT ATIS system and other software assistance. We would also like to thank all the people in AT&T Bell Laboratories Center 1122 who provided technical support and who served as experimenters and transcribers.

REFERENCES

1. Polifroni, J., S. Seneff, V. W. Zue, and L. Hirschman, "ATIS Data Collection at MIT," DARPA SLS Note 8, Spoken Language Systems Group, MIT Laboratory for Computer Science, Cambridge, MA, November, 1990.
2. Polifroni, J., S. Seneff, and V. W. Zue, "Collection of Spontaneous Speech for the ATIS Domain and Comparative Analyses of Data Collected at MIT and TI," *Proc. Fourth DARPA Speech and Language Workshop*, P. Price (ed.), Morgan Kaufmann, February 1991.
3. Seneff, S., L. Hirschman, and V. Zue, "Interactive Problem Solving and Dialogue in the ATIS Domain," *Proc. Fourth DARPA Speech and Language Workshop*, P. Price (ed.), Morgan Kaufmann, February 1991.