THE CARNEGIE MELLON COMMUNICATOR CORPUS

Christina Bennett and Alexander I. Rudnicky

School of Computer Science, Carnegie Mellon University, Pittsburgh PA 15213, USA
{cbennett,air}@cs.cmu.edu

ABSTRACT

As part of the DARPA Communicator program, Carnegie Mellon has, over the past three years, collected a large corpus of speech produced by callers to its Travel Planning system. To date, a total of 180,605 utterances (90.9 hours) have been collected. The data were used for a number of purposes, including acoustic and language modeling and the development of a spoken dialog system. The collection, transcription and annotation of these data prompted us to develop a number of procedures for managing the transcription process and for ensuring accuracy. We describe these, as well as some results based on these data. A portion of this corpus, covering the years 1999-2001, is being published for research purposes.

1. INTRODUCTION

Corpora of human-computer speech are not commonly available. The DARPA Communicator program has created an opportunity to accumulate a large amount of such data for a specific application – travel planning. The corpus is unique in that it is a large, publicly available corpus of human-computer speech.

The assembly of this corpus necessitated the development of procedures that are not often required in a research-based environment. In this paper we describe the collection process as well as the procedures we developed to manage the corpus and to maintain an acceptable level of quality.

2. THE CMU COMMUNICATOR

The CMU Communicator is an advanced spoken dialog system, operating in the domain of travel planning. The domain implemented includes access to flight information (for about 500 destinations world-wide) as well as hotel information (for domestic destinations) and car rentals. Details of the system have been provided in a number of publications, see particularly [1] and [2], and this system, along with others, has been formally evaluated on two occasions. Characteristics that might influence the nature of the speech in this corpus include support for natural language input, barge-in and mixed-initiative dialog. No restrictions were placed on type of telephone (various land-line, cell phone and speakerphone) or individuals calling (including non-native speakers).

3. DATA COLLECTION

All data were collected using the CMU Communicator system, over the course of its development. Since the system was under development, the nature of the speech varied over time; for example, during the earlier period most of the callers were internal – developers and other testers. In August 1999, the system went public and was made available "24x7" through a toll-free number (877-CMU-PLAN). Unlike other systems in the Communicator program, the CMU system did not require pre-registration and would interact with any caller. (Nevertheless, the system did ask callers if they were registered, and if not, to provide their names.) New callers were given the opportunity to listen to a short description of the system's capabilities and received some hints on how to interact with the system. The full text of the introduction is shown in Table 1.

> *You may interrupt these instructions at any time by saying, good enough. The Communicator is a travel planning system with up to the minute flight information. It knows about major U.S. cities, and some international destinations. Here are some tips for a smooth interaction. Please speak clearly and naturally. Do not speak too quickly or too slowly. You can interrupt the system at any time by saying anything you wish. If you need to make a correction, just restate the new information. For example, if you'd like to depart from Cleveland instead of Pittsburgh, you can simply say, I'm departing from Cleveland. Now, here are some keywords you can use. To erase everything so far and start from the beginning, say "start over". To hear the last system response again, say "repeat". To end the call, say "good bye". This is the end of the instructions. If you need help at any time, please say, "help".*

Table 1: Introduction to the CMU Communicator system.

The existence of the system was publicized through a website (http://www.speech.cs.cmu.edu/Communicator) and through other channels, including the distribution of business cards listing the (toll-free) access number for the system. It is not possible to accurately estimate the number of different callers who actually used the system (some did not volunteer their names, others used names that were clearly false). A total of 2066 distinct names (in 12,301 separate instances) were volunteered to the system; we believe that this does roughly approximate the number of distinct speakers in the corpus.

Caller speech (that is, their telephone call) was passed through a Gentner DH20 echo-cancellation device, and then routed into the built-in soundboard in the computer hosting the system. Sampling was at 8kHz and 16 bits. Over the course of the project, 11 different computers hosted the system; however, the bulk of the data was obtained through three computers connected to the toll-free number at various times.

4. TRANSCRIPTION CONVENTIONS

We developed a set of transcription conventions that were focused on the needs of dialog system development. That is, we

needed accurate word-level transcription for purposes of acoustic and language modeling as well as grammar development; in addition, we needed to label other events of importance in acoustic modeling. These included environmental noises, human noises (including fillers), system intrusions (say on echo-cancellation failure) and meta-tags to flag asides and comments. A transcription document was prepared and used for reference and for training transcribers.

The full transcription document is available at (http://www.speech.cs.cmu.edu/Communicator/Corpus). An example of a transcribed session is shown in Table 2.

Hello. Welcome to the Carnegie Mellon Communicator. Are you a registered user?
HI THIS IS <omitted> I'D LIKE TO FLY TO SAN FRANCISCO ON FRIDAY
U: hi this is <omitted> i'd like to fly to San_Francisco on Friday
Hello, <omitted>, it's nice to hear from you again. You can say help at any time. traveling to san francisco on Friday, June sixteenth. [STALL] and what time would you like to depart pittsburgh
IN THE AFTERNOON
U: in the afternoon
a flight in the afternoon. [STALL] okay i have a nonstop on US Airways departing Pittsburgh at two thirty five p.m. arrive San Francisco at four fifty three p.m. Do you want to take that?
YEAH THAT'S FINE
U: yeah that's fine
Will you return to pittsburgh from san francisco?
YEAH I'D LIKE TO COME BACK ON MONDAY
U: yeah i'd like to come back on Monday
traveling on Monday, June nineteenth. [STALL] okay and what time would you like to depart san francisco
IN THE MORNING
U: #begin_feed# in the morning #end_feed#
a flight in the morning. [STALL] i have a US Airways flight departing San Francisco at nine a.m. , arrives Pittsburgh at four thirty six p.m.. Is that OK?
STILL LATER FLIGHT
U: /feed/ is there a later flight
[STALL] the flight after that departs San Francisco at eleven p.m. , arrives Pittsburgh at six forty a.m. Is that OK?
YEAH
U: /feed/ yeah #noise# /hangup/

Table 2: Sample of a transcribed session. [STALL] indicates a variable-content message meant to reassure the caller that the system is "busy" attending to their request and not dead. /feed/ indicates audible system speech. Uppercase lines are decoder output; U: transcription; system prompts are in italics.

5. TRANSCRIPTION TOOLS AND PROCESS

A tool, Scribe, was developed to help transcribers do their work efficiently. The tool allows playback of utterances and manages sessions as created by the Communicator system. To ensure consistency in transcription, all non-lexical items are displayed in panels and can be inserted by double clicks. To ensure spelling consistency, transcriber entries are automatically checked against a domain dictionary and discrepancies are highlighted (though a seemingly misspelled word can be left in by the transcriber).

All utterances were transcribed by one of the transcribers, and then checked by another of the transcribers. The tool logs all transcription and checking activity, allowing overall statistics to be collected on the transcription process.

6. TRANSCRIPTION ACCURACY

We performed an experiment in which we checked accuracy of our transcribe/check process. We sampled material at six-month intervals in the collection and had a panel of three judges listen to each utterance for comparison to the (checked) transcription. A total of 4940 tokens from 1181 utterances, over 99 sessions, were examined. Overall, we determined that 0.83% (±0.25)[1] of the transcribed tokens had errors. The errors were categorized by type, as shown in Figure 1 below.

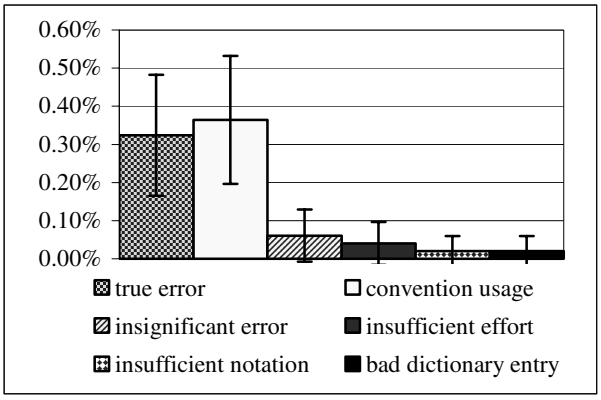

Figure 1: Percentage of transcription errors in the corpus, categorized by type.

6.1. Error categories

The following is an explanation of the seven error categories shown in Figure 1.

The *true error* category contains those errors that we felt were actual mistakes in the human transcription and were not corrected by the second transcriber. For instance, if a word was misspelled or deleted, it would be classified as a true error.

Convention usage represents a perceived miscommunication to the human transcribers regarding the transcription conventions. In particular, we found there to have been some confusion regarding two types of noise, thus the transcribers occasionally used the wrong noise notation.

Some errors were believed to be *insignificant*, that is to say very minor and with no foreseeable repercussions. For example, if a phrase was transcribed as "<Sunday> Sunday evening"

[1] 95% confidence interval

indicating that the first word was a false start, yet the panel disagreed, we felt this was insignificant but an error nonetheless.

If the panel disagreed with a transcription indicating that part of the utterance was spoken but incomprehensible (indicated by /mumble/) (i.e. the spoken word was deemed understandable by the panel), this was viewed as a lack of sufficient effort on the part of the transcriber. These errors were labeled *insufficient effort*.

There was one case found where the panel could not agree on any feasible transcription, given the established conventions; we were then forced to create an *insufficient notation* category.

Another case contained a legitimate spelling error, but since the misspelling was present in the system dictionary, it was not caught by the spell check function of the transcription tool. This error was therefore called *bad dictionary entry*.

6.2. Lexical versus non-lexical errors

An additional analysis of the errors was done to distinguish those that affected lexical items from those that were non-lexical. Lexical tokens are those words that were spoken by the user, as opposed to sounds or other transcription notations (i.e. non-lexical tokens). Figure 2 shows a further breakdown of the errors into categories of lexical and non-lexical errors.

The two bins together represent the errors on all tokens within the transcripts. The first shows the percentage of errors on lexical items in the transcript, whereas the second shows the non-lexical errors. Thus, the combination of these two bins gives the same information as Figure 1.

As can be seen below, transcribers appear to have found it difficult to consistently label non-lexical events. Since the transcription guide emphasized accuracy in lexical transcription, this is perhaps not surprising.

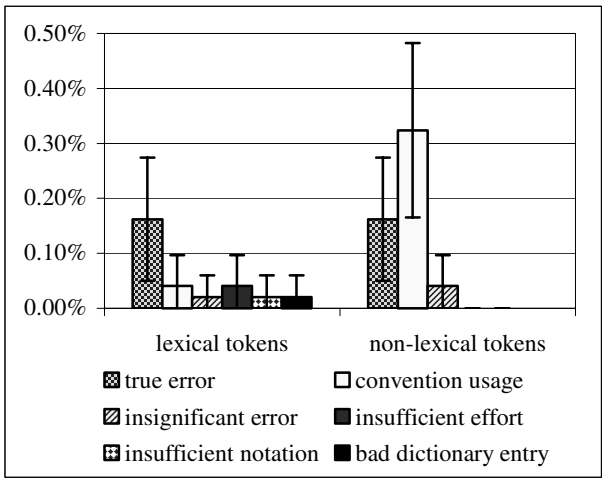

Figure 2: Percentage of transcription errors affecting lexical items versus non-lexical items, further divided by error type.

6.3. Discussion

This very low error rate leads us to believe that we succeeded in producing high-quality transcriptions for the Communicator corpus. We note however that reporting transcription error rate is not consistently a part of corpus documentation, so it is difficult to compare our work to that of others. Exceptions are the SpeechDat corpora (catalogued at http://www.elda.fr) which do specify a transcription error rate. Details of the validation procedure are available (from SPEX, at http://www.spex.nl/validationcentre/) and propose a validation criterion of 5% (utterance) transcription error. Using their computation (though using our sampling approach) yields an error rate of 3.2% (±1.006).

Transcription error, of course, will vary with the level of detail being sought. The inherent difficulty of phonetic transcription, for example, may be reasonably seen to result in higher error than the word-level transcriptions produced for the current corpus. Given our own experience we believe that the informativeness of reporting transcription accuracy justified the relative low cost of producing it.

Overall, we feel that the transcription and auditing procedures we developed for the Communicator project have produced a corpus of speech of high quality that is suitable for research purposes.

7. CORPUS CHARACTERISTICS

The Carnegie Mellon Communicator corpus was collected over a 3 ½ year period, and collection is ongoing as the on-line system will be available indefinitely. Table 3 shows the overall characteristics of the data.

	sessions	utterances	Speech (hours)	Utts per session
All data	15,481	180,605	90.89	11.67
post 1/1/99	11,010	117,695	58.46	10.69

Table 3: Some corpus characteristics.

An analysis performed by Chotimongkol [personal communication] indicated that the language seen by the system changed substantially by the end of 1998 (as measured by language model perplexity). These changes do not however appear to affect the basic characteristics of the corpus. Table 4 lists the 20 most frequent words in the (transcribed) corpus (a total of 468,885 word tokens were transcribed, excluding fragments and false starts). A total of 10,714 distinct words were observed in the corpus. In addition to task-directed speech, we received 2315 comments (at the end of every session users were invited to make comments). Comments were on the average 16.2 words long, while other utterances were on average 2.5 words long (2.4 if we exclude utterances with extraneous remarks, "asides", of which there were 1764).

Word	*Freq (%)*	*Word*	*Freq (%)*
Yes	5.1%	In	1.2%
To	4.9%	I'd	1.1%
No	4.5%	Morning	1.1%
I	3.6%	You	1.0%
The	2.7%	That's	1.0%
Is	1.6%	Tomorrow	1.0%
Go	1.5%	Flight	0.9%
Like	1.4%	From	0.9%
On	1.3%	That	0.9%
A	1.3%	Wanna	0.8%

Table 4: The 20 most frequent words in the corpus.

8. ADDITIONAL ANNOTATION

For certain experiments, we further annotated portions of the corpus for understanding and dialog-level information. This was done manually, and eventually using a web-based tool that allowed the annotator to read through a session gloss and add annotations using a check-mark scheme.

The manual annotations were done on a whole utterance level, whereas the later web-based annotation scheme allowed the annotator to evaluate each portion of the utterance (that is, indicate understanding on a concept-by-concept basis). We obtained whole utterance manual annotations for the period of October through November 1999 and partial utterance annotations for the period of mid June through mid August of 2001. (See [3] for further information on the annotation scheme.) The resulting sub-corpora were used to locate problem regions in dialog and to drive learning-based experiments (e.g., [4]).

9. DIALOG ERROR ANALYSIS

As part of system development, we undertook, at various times, systematic coding and analysis of system performance. A total of 441 dialogs were analyzed to determine the nature and source of errors occurring in calls, based on an expansion of the scheme developed by Constantinides and Rudnicky [5].

Figure 3 shows a chart of system errors for the period of August 1999 through October 2000. Of 8,652 turns analyzed, 35.6% had some type of error. As can be seen the major source of error is recognition (59%), followed by problems with the dialog (12%) and the output (11%).

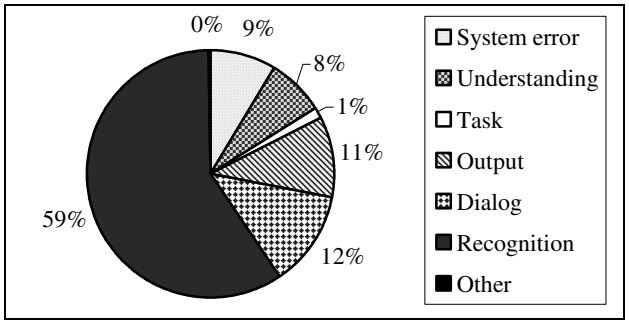

Figure 3: Percentage of system errors attributed to prescribed error types.

10. DISCUSSION

We found that the following characteristics of our effort contributed to its success: a careful two-stage (yet resource-efficient) transcription plus checking process, an emphasis on tools tuned to the task at hand, and easy to use and informative visualization and auditing tools.

As mentioned in section 6, the process of evaluating the quality of our transcriptions did reveal some limitations of our techniques. In particular, when transcribers were uncertain about convention usage, there was no formal way to resolve the confusion. Since we did not actively adapt our transcription conventions with feedback from the transcribers, we did not discover a need for extra notation to denote unresolved uncertainty about an event in an audio file until late in the project.

11. CONCLUSIONS

The Carnegie Mellon Communicator Corpus represents a significant achievement in the development of a large, freely available corpus of human-machine interaction. The assessed quality of its companion transcripts and the availability of associated corpus tools, such as the transcription tool, make it a unique resource.

Further information on obtaining the corpus and its associated tools can be found at http://www.speech.cs.cmu.edu/Communicator/Corpus/.

12. ACKNOWLEDGEMENTS

This research was sponsored in part by the Space and Naval Warfare Systems Center, San Diego, under Grant No. N66001-99-1-8905. The content of the information in this publication does not necessarily reflect the position or the policy of the US Government, and no official endorsement should be inferred.

We would like to thank the following people for their contributions: Karin Gregory, Melissa Schmelzer, Jwan Allen, and Helen Ross for transcription work, Maria Calais Pedro for transcription and participation as a panelist, Paul Constantinides for development of the Scribe transcription tool, and Tania Leibowitz for transcription, annotation, and further development of Scribe.

We would also like to thank Maxine Eskenazi for her numerous contributions to the Communicator data collection effort.

13. REFERENCES

[1] Rudnicky, A. I., Thayer, E., Constantinides, P., Tchou, C., Shern, R., Lenzo, K., Xu, W., Oh, A., "Creating Natural Dialogs in the Carnegie Mellon University Communicator System", *Eurospeech 1999*, V.4: 1531-1534, 1999.

[2] Eskenazi, M., Rudnicky, A., Gregory, K., Constantinides, P., Brennan, R., Bennett, C., Allen, J., "Data Collection and Processing in the Carnegie Mellon Communicator", *Eurospeech 1999*, V.6: 2695-2698, 1999.

[3] Carpenter, P., Jin, C., Wilson, D., Zhang, R., Bohus, D., Rudnicky, A., "Is this Conversation on Track?", *Eurospeech 2001*, pp. 2121-2124, 2001.

[4] Bohus, D. and Rudnicky, A., "Modeling the Cost of Misunderstanding Errors in the CMU Communicator Dialog System", *ASRU 2001* paper a01db097, 2001.

[5] Constantinides, P. C. and Rudnicky, A. I., "Dialog Analysis in the Carnegie Mellon Communicator", *Eurospeech 1999,* V.1: 243-246, 1999.