CORPUS COLLECTION FOR ATIS

Jared Bernstein

SRI International
Menlo Park, CA 94025

PROJECT GOALS

The project goal is to collect and deliver a corpus of speech data that supports DARPA SLS system development. As of February 1991, SRI has set up a hardware and software environment for the collection of spoken interactions with a simulated Air Travel Information System (ATIS), established a data collection procedure, collected and distributed prototype data, and evaluated the prototype data with feedback from the SLS system developers. Having implemented revisions in the environment and procedures, SRI has begun collecting and distributing a corpus of data for ATIS SLS development.

RECENT RESULTS

- Completed a plan for the interface to the relational database, the collection of the prototype and production data, and the subject environment.
- Collected 10 prototype subject sessions, prepared speech and auxiliary files, and shipped data to NIST for distribution to interested SLS developers. Interacted with NIST and with SLS sites to refine certain aspects of the data collection environment and procedures.
- Modified and augmented the tools used in data collection and file preparation; e.g., automated parts of the transcription task and the derivation of additional auxiliary files, and augmented wizard tools to accelerate database responses.
- Provided yield and cost estimates for revised transcription protocols and for extended categorization of utterances.
- Shipped 35 subject sessions to NIST. Recorded and transcribed sessions, generated auxiliary files, prepared session logs and categorized utterances; checked and prepared material for shipment to NIST.
- Shipped 32 more subject sessions to NIST. Categorized, prepared auxiliary files for, checked, and shipped 32 sessions previously recorded and transcribed in summer 1990 under SRI's ATIS SLS contract.

PLANS FOR THE COMING YEAR

- Resume and accelerate data collection in the ATIS domain.
- Document systems and procedures in preparation for export of the wizard data collection system.
- Work with NIST and the DARPA community to define and implement new speech corpus collections.