# NIST-DARPA Interagency Agreement: Spoken Language Program

*David S. Pallett, Principal Investigator*

National Institute of Standards and Technology
Room A-216, Building 225 (Technology)
Gaithersburg, MD

## PROJECT GOALS

1. To coordinate the design, development and distribution of speech and natural language corpora for the DARPA Spoken Language research community.

2. To design, coordinate implementation, and analyze results, of performance assessment "benchmark tests" for DARPA's speech recognition and spoken language understanding systems.

## RECENT RESULTS

1. Completed production of the six-CD-ROM-set for ATIS0, and made this available through the National Technical Information Service.

2. Analyzed the results of the February 1991 Benchmark Tests.

3. Participated in the DARPA MADCOW Group and CCCC activities.

4. Screened and processed the ATIS MADCOW data for production in recordable CD-ROM media (by MIT/LCS), distributed the resultant CD- ROMs, and maintained "on-line" MADCOW data (e.g., transcriptions and other "ancillary files") for access via anonymous ftp.

5. Prepared for and implemented the October 1991 ATIS "dry run" Benchmark Tests with MADCOW data and developed specialized MADCOW Benchmark Test report generation software.

6. Prepared for and implemented the February 1992 ATIS and CSR Benchmark Tests.

7. Screened and processed the Pilot CSR Corpus data for production on recordable CD-ROM media (by MIT/LCS), and distributed the resultant CD-ROMs.

8. Acquired and installed hardware and software to permit CSR corpus collection and transcription at NIST, using MIT/LCS- developed software.

9. Implemented the February 1992 "dry run" Benchmark Test for the CSR Pilot Corpus.

10. In collaboration with Texas Instruments (and using NIST funding) prepared the TI-46 Word Isolated Word Speaker Dependent Speech Database for production on CD-ROM media.

## PLANS FOR THE COMING YEAR

1. Acquire and install hardware and software to run the MIT/LCS "TINA" ATIS system for data collection at NIST for future phases of the MADCOW effort.

2. Acquire and install hardware and software to run the SRI ATIS system for data collection at NIST for future phases of the MADCOW effort.

3. Revise the NIST speech recognition scoring software to more readily accommodate corpora such as the CSR corpus and to provide more informative diagnostics.

4. Develop a comprehensive speech data screening and quality assurance software package.

5. Acquire and make use of recordable CD-ROM technology for limited distribution of speech corpora.

6. Collect, transcribe and annotate (as appropriate) a limited amount of training and test data for the ATIS and CSR corpora.

7. Prepare for and implement Benchmark Tests for the RM, ATIS, and CSR domains (in the late Summer - early Fall 1992 time frame) as required by the DARPA Program Manager and Coordinating Committee.