1. INTRODUCTION

This paper documents benchmark tests implemented within the DARPA Spoken Language Program during the period November, 1992 - January, 1993. Tests were conducted using the Wall Street Journal-based Continuous Speech Recognition (WSJ-CSR) corpus and the Air Travel Information System (ATIS) corpus collected by the Multi-site ATIS Data Collection Working (MADCOW) Group. The WSJ-CSR tests consist of tests of large vocabulary (lexicons of 5,000 to more than 20,000 words) continuous speech recognition systems. The ATIS tests consist of tests of (1) ATIS-domain spontaneous speech (lexicons typically less than 2,000 words), (2) natural language understanding, and (3) spoken language understanding. These tests were reported on and discussed in detail at the Spoken Language Systems Technology Workshop held at the Massachusetts Institute of Technology, January 20-22, 1993.

Tests implemented during this period also included experimental or “dry run” implementation of two new tests. In the WSJ-CSR domain, a “stress test” was implemented, using test material that was drawn from unidentified sub-corpora. In the ATIS domain, an experimental “end-to-end” evaluation was conducted that included examination of the subject-session “logfile”. Following precedents established previously, the results of these dry-run tests are not included as part of the “official” NIST test results and are not discussed at length in this paper.

System descriptions were submitted to NIST by the benchmark test participants and distributed at the Spoken Language Systems Technology Workshop. Additional information describing these systems can be found references 5-23. Detailed information is not available (in published papers) for some systems.

2. WSJ-CSR TESTS: NEW CONDITIONS

2.1. Stress Test

The established benchmark test protocols for speech recognition systems are such that system developers have prior knowledge of the nature of the test material, based on access to similar development test sets. Some developers have consistently declined to report results for material of particular interest to DARPA program management (e.g., for secondary microphone data). Concern has been expressed that the sensitivity or “robustness” of some DARPA-sponsored recognition algorithms has not been adequately probed or the systems “stressed”.

DARPA program management requested that NIST implement, in early December, 1992, a “dry run” of a “stress test” in which the nature of the test material was unspecified. Participating DARPA contractors were required to document and freeze the system configuration used to process the test material prior to implementing the test, and to provide data for a baseline test of this system using the 20K NVP test subset of the Nov ’92 test material, as well as for the stress test set. Test hypotheses were scored by NIST using “conditional scoring” -- partitioning and reporting test results for individual test subsets.

The stress test material consisted of a set of 320 utterance files, chosen from three components: (1) read 20K sentences, for 4 female speakers, (2) read 5K sentences, for 4 female speakers, and (3) spontaneously dictated news articles, for 2 male and 2 female speakers. The read speech included both primary and secondary microphones, so that there were 5 test subsets in all, each consisting of either 60 or 80 utterances.
Reactions to the stress test, as well as to the test results, were mixed. In general, as would be expected, systems with trigram language models did better than those with bigrams. Degradations in performance for the secondary microphone data were relatively smaller for some systems than others -- particularly for those sites that had devoted special effort to the issue of “noise robustness”. However, because the individual test subsets and the number of speakers were small, the results of many of the paired comparison significance tests were inconclusive, suggesting that future applications of such a test procedure must involve larger test subsets.

2.2. New Significance Tests

For several years, NIST has implemented two tests of statistical significance for the results of benchmark tests of speech recognition systems: the McNemar sentence error test (MN) and a Matched-Pair-Sentence-Segment-Word-Error (MAPSSWE) test, on the word error rate found in sentence segments. In more recent tests, NIST has also implemented two additional tests: a Signed-pair (SI) test, and the Wilcoxon signed rank (WI) test. These additional tests are relevant to the word error rates found for individual speakers, and as such are particularly sensitive to the number of speakers in the test set. References to these tests can be found in the literature on nonparametric or distribution-free statistics.

2.3. Uncertainty of Performance Measurement Results

Increasing attention is being paid, at NIST, to evaluating and expressing the uncertainty of measurement results. This attention is motivated, in part, by the realization that “in general, it is not possible to know in detail all of the uses to which a particular NIST measurement result will be put.”[3] Current NIST policy is that “all NIST measurement results are to be accompanied by quantitative measurements of uncertainty”. In substance, the recommended approach to expressing measurement uncertainty is that recommended by the International Committee for Weights and Measures (CIPM).

The CIPM-recommended approach includes: (1) determining and reporting the “standard uncertainty” or positive square root of the estimated variance for each component of uncertainty that contributes to the uncertainty of the measurement result, (2) combining the individual standard uncertainties into a determination of the “combined standard uncertainty”, (3) multiplying the combined standard uncertainty by a factor of 2 (a “coverage factor”, that for normally distributed data corresponds to the 95% confidence interval), and specifying this quantity as the “expanded uncertainty”. The expanded uncertainty, along with the coverage factor, or else the combined standard uncertainty, is to be reported.

The paired-comparison significance tests outlined in the previous section represent specific instantiations of tests that evaluate the validity of null hypotheses regarding differences (in measured performance) between systems. In many cases, however, sufficiently detailed data is not available to implement these tests. In these cases it is important to refer to explicit estimates of uncertainties.

The case of evaluating the uncertainties associated with performance measurements for spoken language technology is particularly complex because of the number of known complicating factors. These factors include properties of the speaker population (e.g., gender, dialect region, speaking rate, vocal effort, etc.), properties of the training and test sets (e.g., vocabulary size, syntactic and semantic properties, microphone/channel, etc.) and other factors [4].

Performance measures used to date within the DARPA spoken language research community (and included in this paper) do not conform to the recommended approach, since the scoring software, in general, generates a single measurement for the ensemble of test data (e.g., one datum indicating word or utterance error rate for the entire multi-speaker, multi-utterance, test subset, rather than the mean error rate for the ensemble of speakers). These single-measurement performance evaluation procedures do not yield estimates of the variances “for each component of uncertainty that contributes to the uncertainty of the measurement result” that are required in order to implement the CIPM-recommended practice.

In future tests, revisions to the scoring software that would permit estimates of the variance across the speaker population (at the least) are in order. However, it would seem to be the case that identifying and obtaining quantitative estimates of “each component of uncertainty that contributes to the uncertainty of the measurement” will be difficult.

3. WSJ-CSR NOVEMBER 1992 TEST MATERIAL

The test material, as distributed, included a total of 16 identified test subsets. In general, these can be sub-categorized five ways: speaker dependent/independent (SD/SI), 5K/20K reference vocabularies, the use of verbalized/non-verbalized punctuation (VP/NVP), read/spontaneous speech, and primary (Sennheiser, close-talking)/secondary microphone. No one participant reported results on all subsets -- most reported results on only one or two, corresponding to conditions of particular local interest and/or algorithmic strength.

All of the test material was drawn from the WSJ-CSR Pilot Corpus that was collected at MIT/LCS, SRI International, and TI. The “spontaneous dictation” data was collected only at SRI.
Individual test set sizes varied from 72 utterances to (more typically) approximately 320 utterances. The number of speakers in each subset varied from 3 to 12 speakers. The actual number of sentence utterances per speaker varied somewhat, because the material was selected in paragraph blocks. A total of 8 secondary microphones was included in the various test subsets, including one speakerphone, a telephone handset, 3 boundary effect microphones (Crown PCC-160, PZM-6FS, and Shure SM91), two lavalier microphones, and a desk-stand mounted microphone.

4. WSJ-CSR TEST PROTOCOLS

Test protocols were similar to prior speech recognition benchmark tests. Test material was shipped to the participating sites on October 20th, results were reported on Nov. 23rd, and NIST reported scored results via ftp to the participants on Dec. 2nd. The stress test was conducted between Nov. 30th and Dec. 15th.

A “required baseline” test was defined for all participants. It consisted of processing the 5K word speaker independent, non-verbalized punctuation test set using a (common) bigram grammar. Six sites reported 5K baseline test results.

5. WSJ-CSR TEST SCORING

As for the test protocols, much of the scoring was routine, except for one new additional factor. Since previous “official” CSR benchmark tests had not included spontaneous speech, the community had not reviewed the adequacy of the transcription convention used for spontaneous speech, and several inconsistencies in the transcriptions were noted following release of the preliminary results. Some of these inconsistencies were resolved prior to releasing “official” results.

6. WSJ-CSR TEST PARTICIPANTS

Participants in these WSJ-CSR tests included the following DARPA contractors: BBN, CMU, Dragon Systems, MIT Lincoln Laboratory, and SRI International. A “volunteer” participant was the French CNRS LIMSI. LIMSI declined to participate in the “stress test”.

7. WSJ-CSR BENCHMARK TEST RESULTS AND DISCUSSION

7.1. Test Results: Word and Utterance (Sentence) Error Rates

Table 1 presents the results for the several test sets on which results were reported. Section I of that table includes results reported by Paul at MIT Lincoln Laboratory [5] for Longitudinal Speaker Dependent (LSD) technology. Section II includes results reported by BBN for Speaker Dependent (SD) technology. Section III includes the results of Speaker Independent (SI) technology, for a number of sites for (a) the 20K NVP test set for both baseline and non-baseline SI systems, (b) the 5K NVP test set for both baseline and non-baseline SI systems, (c) the 5K NVP test set “other microphone” test set data, and (d) the 5K VP test set (on which only LIMSI reported results [6]). Section IV of Table 1 includes the results reported by BBN for the Spontaneous Dictation test set.

For the test set on which the largest number of results were reported -- the 5K NVP set, using the close-talking microphone -- the lowest word error rates were reported by CMU [7-9]: 6.9% for the baseline, bigram language model, and 5.3% using a trigram language model. The range of word error rates for the baseline condition for all systems tested was 6.9% to 15.0%, while for non-baseline conditions, the range was from 5.3% to 16.8%.

For the 5K NVP test set's secondary microphone data, as reported by CMU [8] and SRI [10,11], word error rates ranged from 17.7% to 38.9%.

For the 20K NVP test set, on which other baseline data were reported, the word error rates range from 15.2% to 27.8%.

The lowest error rate, reported by CMU, can be shown to be significantly different for all 4 significance tests when compared with the Dragon [13] and MIT Lincoln systems, but shown to be significantly different only for the MAPSSWE test when compared with the BBN system [14]. Thus the performance differences between the CMU and BBN systems for this baseline condition test are very small.

7.2. Significance Test Results

Table 2 presents the results, in a matrix form, of 4 paired-comparison significance tests for the baseline tests for the 5K NVP test set. The convention in this form of results tabulation is that if the result of a null-hypothesis test is valid, the word “same” is printed in the appropriate matrix element. If the null hypothesis is not valid, the identifier for the system with the lower (and significantly different) error rate is printed.

For this test set, recall that the CMU system (here identified as cmul-a) had a word error rate of 6.9%. By comparing the results for the CMU system with the other 5 systems reporting baseline results, note that the significance tests results all indicate that the null hypothesis is not valid. In other words, the error rates for the CMU system are significantly different (lower) than those for the other 5 systems for this test set and baseline conditions.

In general, for this test set, with 12 speakers and 310 utterances, the Wilcoxon signed rank test (WI) is more sensitive than the (ordinary) sign test (SI). As noted in previous tests, the McNemar test (MN), operating on the sentence error rate, is in general less sensitive than the matched-pair-sentence segment word error rate test (MAPSSWE).
8. ATIS TESTS: NEW CONDITIONS

Within the community of ATIS system developers, there is a continuing search for evaluation methodologies to complement the current evaluation methodology. In particular there is a recognized need for evaluation methodologies that can be shown to correlate well with expected performance of the technology in applications. Toward the end of 1992, several sites participated in an experimental “end-to-end” evaluation to assess systems in an interactive form. The end-to-end evaluation included (1) objective measures such as timing information and time to task completion, (2) human-derived judgements on correctness of system answers and user solutions, and (3) a user satisfaction questionnaire. The results of this “dry run” complementary evaluation experiment are reported by Hirschman et al. in [15].

9. ATIS TEST MATERIAL

Test material for the ATIS benchmark tests consisted of 1002 queries, for 118 subject-scenarios, involving 37 subjects. It was selected by NIST from set-aside material drawn from data previously collected within the MAD-COW community at AT&T, BBN, CMU, MIT/LCS, and SRI. The selection and composition of this test material is described in more detail in [15].

As in previous years, queries were categorized into two categories of “answerable” queries, Class A, which are context-independent, and Class D, which are context-dependent; and “unanswerable”, or Class X queries. In the final adjudicated test set, there were a total of 427 Class A queries, 247 Class D queries, and 328 Class X queries.

10. ATIS TEST PROTOCOLS

As was the case for the speech recognition benchmark tests, ATIS test protocols were similar to prior ATIS benchmark tests. The test material was shipped to the participating sites on October 20th, results were reported on Nov. 16th, and NIST reported preliminary scored results via ftp to the participants on Nov. 20th. After the process of formal “adjudication” had taken place, official results were reported on Dec. 20th.

11. ATIS SCORING AND ADJUDICATION

After the preliminary scoring results were distributed, the participating sites were invited to send requests for adjudication (“bug reports”) to NIST, asking for changes in the scoring of specific queries. A total of 146 of these bug reports were adjudicated by NIST and SRI jointly. Since many of these requests for adjudication were duplicates, the number of distinct problems reported was less than 100. A decision was made on each request for adjudication and the corrected reference material or procedure was used in a final adjudicated re-run of the evaluation. The judgment was in favor of the plaintiff in approximately 2/3 of the cases.

A number of problems uncovered by this procedure were systematic, in that the same root problem affected several different queries. Most of these were simply human error, which can be made less likely in the future by working less hectically and making software to double-check the test material.

The major problem that cannot be attributed to just human error is that of transcribing and scoring correctly speech that is difficult to hear and understand. Some of this speech was “sotto voce”; some was mispronounced; some was truncated; and in some cases the phonetic transcription would have been unproblematical but division into lexical words was unclear, as in some contractions and compound words. The short-term solution adopted was just to make our best judgement on orthographic transcription, considering both acoustics and higher-level language modeling. But a better long-term cure is to make and use transcriptions that can indicate alternatives when the word spoken is uncertain; proposals to this effect are being considered by relevant committees.

12. ATIS TEST PARTICIPANTS

Participants in these ATIS tests included the following DARPA contractors: BBN, CMU, MIT Laboratory for Computer Science (MIT/LCS), and SRI. There were several “volunteers”: AT&T Bell Laboratories [16], who have participated in previous years; Paramax [17], not a DARPA contractor at the time of these tests, but who have also participated in prior years’ tests; and two participants from Canada, CRIM and INRS. A total of 8 system developers participated in some of the tests (i.e., the NL tests).

13. ATIS BENCHMARK TEST RESULTS

13.1. ATIS SPontaneous speech RECognition Tests (SPREC)

Table 3 presents the results for the SREC tests for all systems and all subsets of the data. For the interesting case of the subset of all answerable queries, Class A+D, the word error rate ranged from 4.3% to 100%. The lowest value was reported by BBN [18,19], and the value of 100% was reported by INRS, for an incomplete ATIS system that (in effect) rejected every utterance, resulting in a scored word deletion error of 100%.

Table 4 presents a matrix tabulation of ATIS SREC results for the set of answerable queries, Class A+D. This form of matrix tabulation is discussed in [2] for the February 1992 test results. Considerable variability can be noted for the performance of some systems on “local data”, and there are indications of varying degree of difficulty for the subsets collected at different sites. As in the Feb.’92 test set, participants noted the presence of more disfluencies in the AT&T data than for other originating sites.
Word error rates for the "volunteers" in these tests (AT&T, CRIM and INRS) are in general higher than for DARPA contractors, perhaps reflecting a reduced level-of-effort, relative to "funded" efforts.

Table 5 presents the results, in a matrix form, of 4 paired-comparison significance tests for the 7 SPREC systems for the Class A+D subset.

For this test set, recall that the BBN system (here identified as bbn2a_d) had a word error rate of 4.3%. By comparing the results for this BBN system with the other 6 ATIS SPREC systems, note that the null hypothesis is not valid for all 4 significance tests for the comparisons with the AT&T, CRIM, INRS, MIT/LCS and SRI systems. In other words, the differences in performance are significant. However, when comparing the BBN and CMU SPREC systems, the null hypothesis is valid for 3 of the 4 tests. Thus, as was the case for the WSJ-CSR data, the performance differences, in this case for ATIS spontaneous speech, between the CMU and BBN speech recognition systems are very small.

13.2. Natural Language Understanding Tests (NL)

Table 6 presents a tabulation of the results for the NL tests for all systems and the "answerable" ATIS queries, Class A+D, as well as the subsets, Class A and Class D.

For the set of answerable queries, Class A+D, the weighted error ranges from 101.5% to 12.3%. For the Class A queries, the range is from 79.9% to 12.2%. And for the Class D queries, the range is from 138.9% to 12.6%. In each case, the lowest weighted error rate was reported by the CMU system [20].

Note that in general performance is considerably worse for Class D than for Class A. However, for the CMU and MIT/LCS [21] systems, performance for the Class D test material is comparable to that for Class A. These systems would appear to have superior procedures for handling context.

Table 7 presents a matrix tabulation of the NL results for the several subsets of test material. Note, however, that since the differences in performance between DARPA-contractor-developed systems and those of "volunteers", in general, are significant, the column averages presented in this table are not very informative.

Of the 3 CRIM systems, the best performing one (crim3) is one using neural networks to classify each query into 1 of 10 classes based on relation names in the underlying ATIS relational database, with subsequent use of specific parsers built for each class and another parser that determines the constraints [22].

There are two SRI NL systems [23]. The SRI NL-TM system, here designated sri1, uses template matching to generate database queries. The other SRI system, termed the "Gemini+TM ATIS System" by SRI, and here designated sri2, is an integration of SRI's unification-based natural-language processing system and the Template Matcher. Differences in performance do not appear to be pronounced.

As in previous ATIS NL tests, it is important to note that appropriate tests of statistical significance have not yet been developed for ATIS NL tests. Small differences in weighted error rate are probably of no significance. However, large, systematic, differences are noteworthy, even if of unknown statistical significance. The weighted error rates for the CMU NL system, which are in many cases approximately one-half those of the next best systems, are certainly noteworthy.

13.3. Spoken Language System Understanding (SLS)

Table 8 presents a tabulation of the results for the SLS tests for all systems and the "answerable" ATIS queries, Class A+D, as well as the subsets, Class A and Class D.

For the set of answerable queries, Class A+D, the weighted error ranges from 100% to 21.6%. For the Class A queries, the range is from 100% to 19.7%. And for the Class D queries, the range is from 140.1% to 23.9%. As in the case of the NL test results, and in each case, the lowest weighted error rate was reported for the CMU system.

The INRS data signify 100% usage of the No_Answer option, since the INRS SPREC system provided null hypothesis strings, causing the NL component to return the No_Answer response.

Note again that the CMU and MIT/LCS systems both handle context sensitivity well.

Table 9 presents a matrix tabulation of the SLS results for the several subsets of test material.

For the ATIS SLS with lowest overall weighted error rate (21.6%), the cmu1 system, there is an almost ten-fold range in error rate over the several test subsets: from 37.1%, for the AT&T subset, to 3.9% for the SRI subset. The CMU SLS weighted error rates for Class A+D are approximately two-thirds those of the next-best-performing systems, although for the Class A subset, differences in performance between the CMU system and the BBN and SRI systems are less pronounced.

14. ACKNOWLEDGEMENT

At NIST, our colleague Nancy Dalgren contributed significantly to the DARPA ATIS community and had a major role in annotating data and implementing "bug fixes" in collaboration with the SRI annotation group and others. Nancy was severely injured in an automobile accident in November, 1992, and is undergoing rehabilitation therapy for treat-
ment of head trauma. It is an understatement to say that we miss her very much.

Brett Tjaden also assisted us at NIST in preparing test material and other ways.

The cooperation of the many participants in the DARPA data and test infrastructure -- typically several individuals at each site -- is gratefully acknowledged.

References

1. Pallett, D.S., “DARPA February 1992 Pilot Corpus CSR ‘Dry Run’ Benchmark Test Results”, in Proceedings of Speech and Natural Language Workshop, February 1992 (M. Marcus, ed.) ISBN 1-55860-272-0, Morgan Kaufmann Publishers, Inc., pp. 382-386.

2. Pallett, D.S., et al., “DARPA February 1992 ATIS Benchmark Test Results”, in Proceedings of Speech and Natural Language Workshop, February 1992 (M. Marcus, ed.) ISBN 1-55860-272-0, Morgan Kaufmann Publishers, Inc., pp. 15-27.

3. Taylor, B.N. and Kuyatt, C.E., “Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results”, NIST Technical Note 1297, January 1993.

4. Pallett, D.S. “Performance Assessment of Automatic Speech Recognizers”, J. Res. National Bureau of Standards, Volume 90, #5, Sept.-Oct. 1983, pp. 371-387.

5. Paul, D.B. and Nencioglu, B.F., “The Lincoln Large-Vocabulary Stack-Decoder HMM CSR”, Proceedings of ICASSP’93.

6. Gauvain, J.L., et al., “LIMSI Nov92 Evaluation”, Oral Presentation at the Spoken Language Systems Technology Workshop, January 20-22, 1993, Cambridge, MA.

7. Huang, X., et al., “The SPHINX-II Speech Recognition System: An Overview”, Computer Speech and Language, in press (1993).

8. Alleva, F., et al., “An Improved Search Algorithm for Continuous Speech Recognition”, Proceedings of ICASSP’93.

9. Hwang, M.Y., et al., “Predicting Unseen Triphones with Senones”, Proceedings of ICASSP’93.

10. Liu, F.-H., et al., “Efficient Cepstral Normalization for Robust Speech Recognition”, in Proceedings of the Human Language Technology Workshop, March 1993 (M. Bates, ed.) Morgan Kaufmann Publishers, Inc.

11. Murveit, H., et al., “Large-Vocabulary Dictation using SRI’s DECIPHER (tm) Speech Recognition System: Progressive Search Techniques”, Proceedings of ICASSP’93.

12. Murveit, H., et al., “Progressive-search Algorithms for Large Vocabulary Speech Recognition”, in Proceedings of the Human Language Technology Workshop, March 1993 (M. Bates, ed.) Morgan Kaufmann Publishers, Inc.

13. Roth, R., et al., “Large Vocabulary Continuous Speech Recognition of Wall Street Journal Data”, Proceedings of ICASSP’93.

14. Schwartz, R., et al., “Comparative Experiments on Large Vocabulary Speech Recognition”, in Proceedings of the Human Language Technology Workshop, March 1993 (M. Bates, ed.) Morgan Kaufmann Publishers, Inc.

15. Hirschman, L., et al., “Multi-Site Data Collection and Evaluation in Spoken Language Understanding”, in Proceedings of the Human Language Technology Workshop, March 1993 (M. Bates, ed.) Morgan Kaufmann Publishers, Inc.

16. Tzoukermann, E., (Untitled) Oral Presentation at the Spoken Language Systems Technology Workshop, January 20-22, 1993, Cambridge, MA.

17. Linebarger, M.C., Norton, L.M. and Dahl, D.A., “A portable approach to last resort parsing and interpretation”, in Proceedings of the Human Language Technology Workshop, March 1993 (M. Bates, ed.) Morgan Kaufmann Publishers, Inc.

18. Bates, M., et al., “Design and Performance of HARC, the BBN Spoken Language Understanding System”, Proceedings of ICSLP-92, Banff, Alberta, Canada, October, 1992.

19. Bates, M., et al., “The BBN/HARC Spoken Language Understanding System”, Proceedings of ICASSP’93.

20. Ward, W. and Issar, S., “CMU ATIS Benchmark Evaluation”, Oral Presentation at the Spoken Language Systems Technology Workshop, January 20-22, 1993, Cambridge, MA.

21. Glass, et al., “The MIT ATIS System: January 1993 Progress Report”, Oral Presentation at the Spoken Language Systems Technology Workshop, January 20-22, 1993, Cambridge, MA.

22. Cardin, R., et al., “CRIM’s Speech Understanding System for the ATIS Task”, Oral Presentation at the Spoken Language Systems Technology Workshop, January 20-22, 1993, Cambridge, MA.

23. Dowding, J., et al., “Gemini: A Natural Language System for Spoken-Language Understanding”, in Proceedings of the Human Language Technology Workshop, March 1993 (M. Bates, ed.) Morgan Kaufmann Publishers, Inc.
I. Longitudinal Speaker Dependent Tests

a. LSD EVL 20K NVP Test Set

| Systems   | W.Err | U.Err | IDENTIFIER                               |
|-----------|-------|-------|------------------------------------------|
| mlit_114-n | 14.6  | 78.2  | LL NOV92 CSR LSD 20K CLOSED NVP BIGRAM   |
| mlit_115-n | 11.2  | 71.8  | LL NOV92 CSR LSD 20K CLOSED NVP TRIGRAM  |

b. LSD EVL 20K VP Test Set

| mlit_114-1 | 11.6  | 70.7  | LL NOV92 CSR LSD 20K CLOSED VP BIGRAM    |
| mlit_115-1 | 7.6   | 56.0  | LL NOV92 CSR LSD 20K CLOSED VP TRIGRAM   |

II. Speaker Dependent Tests

a. SD EVL 5K NVP Test Set

| Systems   | W.Err | U.Err | IDENTIFIER                               |
|-----------|-------|-------|------------------------------------------|
| bbn2-e    | 8.2   | 54.5  | BBN NOV92 CSR BYBLOS SD-600 5K BIGRAM    |
| bbn3-e    | 6.1   | 44.5  | BBN NOV92 CSR BYBLOS SD-600 5K TRIGRAM   |

III. Speaker Independent Tests: Read Speech

a. SI EVL 20K NVP Test Set (Baseline Tests)

| Systems   | W.Err | U.Err | IDENTIFIER                               |
|-----------|-------|-------|------------------------------------------|
| bbn1-d    | 16.7  | 81.1  | BBN NOV92 CSR BYBLOS SI-12 20K BIGRAM BASELINE |
| cmu2-d    | 15.2  | 79.0  | CMU NOV92 CSR SPHINX-II SI-04 20K BASELINE |
| dragon1-d | 25.0  | 86.8  | DRAGON NOV92 CSR MULTIPLE SI-12 20K NVP BASELINE |
| mit_111-d | 25.2  | 88.0  | LL NOV92 CSR SI-84 20K OPEN NVP BIGRAM BASELINE |

b. SI EVL 20K NVP Test Set (Non-Baseline Tests)

| Systems   | W.Err | U.Err | IDENTIFIER                               |
|-----------|-------|-------|------------------------------------------|
| bbn3-d    | 14.8  | 75.7  | BBN NOV92 CSR BYBLOS SI-12 20K TRIGRAM   |
| cmu2-d    | 12.8  | 71.8  | CMU NOV92 CSR SPHINX-II SI-04 20K TRIGRAM |
| dragon1-d | 24.8  | 87.4  | DRAGON NOV92 CSR GD SI-12 20K NVP        |
| mit_111-d | 27.8  | 87.4  | DRAGON NOV92 CSR GI SI-12 20K NVP        |

IV. Speaker Independent Tests: Spontaneous Speech

a. SI SPONTANEOUS DICTATION NVP Test Set

| Systems   | W.Err | U.Err | IDENTIFIER                               |
|-----------|-------|-------|------------------------------------------|
| bbn1-a    | 0.7   | 63.6  | LL NOV92 CSR BYBLOS SI-12 5K NVP BASELINE |
| cmu1-a    | 6.9   | 57.6  | LL NOV92 CSR SPHINX-II SI-04 5K BASELINE |
| dragon3-a | 14.1  | 78.2  | DRAGON NOV92 CSR MULTIPLE SI-12 5K NVP BASELINE |
| mit_111-a | 19.4  | 84.1  | LL NOV92 CSR SI-84 20K OPEN NVP TRIGRAM ADAPTIVE |

b. SI EVL 5K NVP Test Set (Baseline Tests)

| Systems   | W.Err | U.Err | IDENTIFIER                               |
|-----------|-------|-------|------------------------------------------|
| bbn1-a    | 7.3   | 53.0  | BBN NOV92 CSR BYBLOS SI-12 5K TRIGRAM    |
| cmu1-a    | 5.4   | 45.2  | CMU NOV92 CSR SPHINX-II SI-04 5K NVP     |
| cmu2-a    | 8.1   | 63.0  | CMU NOV92 SPHINX-II MFCDCN W/O COMP CSR SI-84 5K NVP |
| cmu3-a    | 9.4   | 67.9  | CMU NOV92 SPHINX-II MFCDCN W/ COMP CSR SI-84 5K NVP |
| cmu4-a    | 8.4   | 63.0  | CMU NOV92 SPHINX-II CDEN W/O COMP CSR SI-84 5K NVP |
| cmu5-a    | 8.1   | 62.2  | CMU NOV92 SPHINX-II CDEN W COMP CSR SI-84 5K NVP |
| dragon1-a | 13.6  | 76.7  | DRAGON NOV92 CSR GD SI-12 5K NVP         |
| mit_112-a | 10.5  | 61.2  | LL NOV92 CSR SI-84 5K CLOSED NVP TRIGRAM |
| mit_113-a | 9.1   | 56.7  | LL NOV92 CSR SI-84 5K CLOSED NVP TRIGRAM ADAPTIVE |

c. SI EVL 5K NVP OTHER MICROPHONE Test Set

| Systems   | W.Err | U.Err | IDENTIFIER                               |
|-----------|-------|-------|------------------------------------------|
| cmu1-c    | 38.5  | 88.2  | CMU NOV92 SPHINX-II MFCDCN W/O COMP CSR SI-84 5K NVP |
| cmu2-c    | 17.7  | 75.8  | CMU NOV92 SPHINX-II MFCDCN W/ COMP CSR SI-84 5K NVP |
| cmu3-c    | 38.9  | 87.3  | CMU NOV92 SPHINX-II CDEN W/O COMP CSR SI-84 5K NVP |
| cmu4-c    | 38.9  | 87.3  | CMU NOV92 SPHINX-II CDEN W/ COMP CSR SI-84 5K NVP |
| cmu5-c    | 19.3  | 77.9  | CMU NOV92 SPHINX-II CDEN W COMP CSR SI-84 5K NVP |
| sril-c    | 27.3  | 87.6  | SRI NOV92 CSR DECIPHER(TM) SI-84 5K NVP BASELINE |

Table I: WSJ-CSR Benchmark Test Results
| Test Name Abbrev. |
|------------------|
| Matched Pair Sentence Segment (Word Error) Test | MP |
| Signed Paired Comparison (Speaker Word Accuracy) Test | SI |
| Wilcoxon Signed Rank (Speaker Word Accuracy) Test | WI |
| McNemar (Sentence Error) Test | MN |

Table 2: Significance Test Results: Baseline Tests Using the 5K NVP Test Set
(See text for explanation of format)
### ATIS SPREC Test Results

#### Class A\+D\+X Subset

| W. Err | Corr | Sub | Del | Ins | U. Err | # Utt. | Description |
|--------|------|-----|-----|-----|-------|--------|-------------|
| att2-adx | 11.7 | 90.8 | 6.8 | 2.4 | 2.5 | 52.4 | 967 | ATT Nov '92 SPREC Results |
| bbn2-adx | 7.6 | 94.2 | 4.2 | 1.6 | 1.8 | 35.6 | 967 | BBN Nov '92 SPREC Results |
| cmu2-adx | 8.3 | 92.9 | 4.2 | 2.9 | 1.2 | 38.3 | 967 | CMU Nov '92 SPREC Results |
| crim4-adx | 19.3 | 84.1 | 12.1 | 3.8 | 3.4 | 64.1 | 967 | CRIM Nov '92 SPREC Results |
| inrs2-adx | 12.6 | 89.8 | 7.3 | 2.9 | 2.4 | 47.8 | 967 | INRS Late Nov '92 SPREC Results |
| sri3-adx | 9.1 | 93.2 | 5.4 | 1.4 | 2.3 | 43.3 | 967 | SRI Nov '92 SPREC Results |

#### Class A\+D Subset

| W. Err | Corr | Sub | Del | Ins | U. Err | # Utt. | Description |
|--------|------|-----|-----|-----|-------|--------|-------------|
| att2-a_d | 8.4 | 93.6 | 4.6 | 1.8 | 2.0 | 44.7 | 674 | ATT Nov '92 SPREC Results Class A\+D |
| bbn2-a_d | 4.3 | 96.7 | 2.5 | 0.9 | 0.9 | 25.2 | 674 | BBN Nov '92 SPREC Results Class A\+D |
| cmu2-a_d | 4.7 | 96.0 | 2.6 | 1.2 | 0.7 | 28.9 | 674 | CMU Nov '92 SPREC Results Class A\+D |
| crim4-a_d | 14.1 | 88.7 | 8.4 | 2.9 | 2.8 | 56.4 | 674 | CRIM Nov '92 SPREC Results Class A\+D |
| inrs2-a_d | 100.0 | 0.0 | 0.0 | 100.0 | 0.0 | 100.0 | 674 | INRS Late Nov '92 SPREC Results Class A\+D |
| mit_lcs2-a_d | 8.1 | 93.3 | 4.5 | 2.2 | 1.4 | 37.8 | 674 | MIT-LCS Nov '92 SPREC Results Class A\+D |
| sri3-a_d | 5.7 | 95.7 | 3.5 | 0.9 | 1.4 | 33.8 | 674 | SRI Nov '92 SPREC Results Class A\+D |

#### Class A Subset

| W. Err | Corr | Sub | Del | Ins | U. Err | # Utt. | Description |
|--------|------|-----|-----|-----|-------|--------|-------------|
| att2-a | 8.0 | 93.8 | 4.4 | 1.8 | 1.8 | 45.4 | 427 | ATT Nov '92 SPREC Results Class A |
| bbn2-a | 4.0 | 96.7 | 2.3 | 1.0 | 0.8 | 25.3 | 427 | BBN Nov '92 SPREC Results Class A |
| cmu2-a | 4.4 | 96.1 | 2.7 | 1.2 | 0.5 | 30.7 | 427 | CMU Nov '92 SPREC Results Class A |
| crim4-a | 13.5 | 88.9 | 8.0 | 3.1 | 2.4 | 57.8 | 427 | CRIM Nov '92 SPREC Results Class A |
| inrs2-a | 100.0 | 0.0 | 0.0 | 100.0 | 0.0 | 100.0 | 427 | INRS Late Nov '92 SPREC Results Class A |
| mit_lcs2-a | 7.8 | 93.5 | 4.4 | 2.2 | 1.3 | 38.2 | 427 | MIT-LCS Nov '92 SPREC Results Class A |
| sri3-a | 5.2 | 96.0 | 3.2 | 0.9 | 1.1 | 34.2 | 427 | SRI Nov '92 SPREC Results Class A |

#### Class D Subset

| W. Err | Corr | Sub | Del | Ins | U. Err | # Utt. | Description |
|--------|------|-----|-----|-----|-------|--------|-------------|
| att2-d | 9.2 | 93.2 | 5.0 | 1.7 | 2.4 | 43.3 | 247 | ATT Nov '92 SPREC Results Class D |
| bbn2-d | 4.8 | 96.5 | 2.8 | 0.7 | 1.3 | 25.1 | 247 | BBN Nov '92 SPREC Results Class D |
| cmu2-d | 5.4 | 95.7 | 3.2 | 1.1 | 1.1 | 25.9 | 247 | CMU Nov '92 SPREC Results Class D |
| crim4-d | 15.4 | 88.2 | 9.4 | 2.4 | 3.6 | 53.8 | 247 | CRIM Nov '92 SPREC Results Class D |
| inrs2-d | 100.0 | 0.0 | 0.0 | 100.0 | 0.0 | 100.0 | 247 | INRS Late Nov '92 SPREC Results Class D |
| mit_lcs2-d | 8.9 | 92.9 | 5.0 | 2.1 | 1.8 | 37.2 | 247 | MIT-LCS Nov '92 SPREC Results Class D |
| sri3-d | 7.1 | 95.0 | 4.1 | 0.8 | 2.1 | 33.2 | 247 | SRI Nov '92 SPREC Results Class D |

#### Class X Subset

| W. Err | Corr | Sub | Del | Ins | U. Err | # Utt. | Description |
|--------|------|-----|-----|-----|-------|--------|-------------|
| att2-x | 18.5 | 85.1 | 11.3 | 3.6 | 3.5 | 70.3 | 293 | ATT Nov '92 SPREC Results Class X |
| bbn2-x | 14.5 | 89.2 | 7.8 | 3.0 | 3.7 | 59.0 | 293 | BBN Nov '92 SPREC Results Class X |
| cmu2-x | 15.6 | 86.6 | 7.0 | 6.5 | 2.2 | 59.7 | 293 | CMU Nov '92 SPREC Results Class X |
| crim4-x | 10.1 | 74.7 | 19.7 | 5.6 | 4.8 | 81.6 | 293 | CRIM Nov '92 SPREC Results Class X |
| inrs2-x | 100.0 | 0.0 | 0.0 | 100.0 | 0.0 | 100.0 | 293 | INRS Late Nov '92 SPREC Results Class X |
| mit_lcs2-x | 21.7 | 82.6 | 12.9 | 4.6 | 4.2 | 70.6 | 293 | MIT-LCS Nov '92 SPREC Results Class X |
| sri3-x | 15.8 | 88.1 | 9.4 | 2.4 | 4.0 | 64.8 | 293 | SRI Nov '92 SPREC Results Class X |

Table 3: ATIS SPREC Benchmark Test Results

15
Nov92 ATIS SPREC Test Results

Matrix column present results for Test Data Subsets collected at several sites and matrix rows present results for different systems.

Numbers printed at the top of the matrix columns indicate the number of utterances in the Test Data (subset) from the corresponding site.

"Overall Totals" (column) present results for the entire Class A+D Subset for the system corresponding to that matrix row.
"Foreign Syst. Totals" present results accumulated for "foreign systems" (i.e., excluding results for the system(s) developed at the site responsible for collection of that Test Data subset.)

Table 4: ATIS SPREC Results: Class (A,D) by Collection Site

| Test Name Abbrev. | Matched Pair Sentence Segment (Word Error) Test | McNemar (Sentence Error) Test | Wilcoxon Signed Rank (Speaker Word Accuracy) Test | Signed Paired Comparison (Speaker Word Accuracy) Test | Overall Totals | Foreign Totals |
|-------------------|-----------------------------------------------|-----------------------------|------------------------------------------------|------------------------------------------------|----------------|----------------|
| att2-a_d          |                                |                             |                                               |                                               | 6.7 ± 2.7      | 6.7 ± 2.7      |
| bbn2-a_d          |                                |                             |                                               |                                               | 6.5 ± 2.4      | 6.5 ± 2.4      |
| cmu2-a_d          |                                |                             |                                               |                                               | 6.4 ± 2.6      | 6.4 ± 2.6      |
| crlm4-a_d         |                                |                             |                                               |                                               | 6.3 ± 2.8      | 6.3 ± 2.8      |
| mit Lov2-a_d      |                                |                             |                                               |                                               | 6.6 ± 2.1      | 6.6 ± 2.1      |
| sr13-a_d          |                                |                             |                                               |                                               | 6.2 ± 2.0      | 6.2 ± 2.0      |

Table 1: Significance Test Results: ATIS SPREC Systems
| Class (A+D) Set | Originating Site of Test Data | Overall | Foreign |
|----------------|-------------------------------|---------|---------|
|                | ATT | BBN | CMU | MIT | SRI | Totals | Coll. Site | Totals |
| att1           | 71  14 | 79  29 | 16 | 93  45 | 4 | 137  25 | 5 | 135  14 | 3 | 515 127 | 32 | 444 113 | 28 |
| bbn1           | 80  16 | 4 | 64  23 | 13 | 65  32 | 3 | 82  15 | 3 | 89  9 | 2 | 76  19 | 5 | 76  19 | 5 |
| cmul           | 36  0.0 | 59.7 | 66.2 | 32.9 | 24.0 | 20.4 | 42.4 | 43.4 |
| crim1          | 76  3 | 10 | 95  15 | 14 | 116  15 | 11 | 150  5 | 12 | 136  9 | 7 | 573  47 | 54 | 478 32 | 40 |
| crim2          | 85  3 | 11 | 77  12 | 11 | 82  11 | 8 | 90  3 | 7 | 89  6 | 5 | 85  7 | 8 | 87  6 | 7 |
| crim3          | 16.0 | 105.9 | 15.1 | 64.6 | 13.2 | 14.4 | 22.0 | 18.9 |
| inrs1          | 84  5 | 0 | 100  20 | 4 | 138  4 | 0 | 158  8 | 1 | 150  2 | 0 | 630  39 | 5 | 492 35 | 5 |
| mit_lcsi       | 94  0 | 0 | 81  16 | 3 | 97  3 | 0 | 95  5 | 1 | 99  1 | 0 | 93  6 | 1 | 92  7 | 1 |
| paramax        | 11.2 | 35.5 | 5.6 | 10.2 | 2.6 | 12.3 | 14.1 |
| srl1           | 82.0 | 77.4 | 89.4 | 56.9 | 50.7 | 69.4 | 69.4 |
| srl2           | 52.8 | 54.8 | 56.3 | 52.7 | 34.2 | 49.7 | 49.7 |
| crim3          | 63  21 | 5 | 88  32 | 4 | 101  39 | 2 | 119  40 | 8 | 126  26 | 0 | 497 158 | 19 | 497 158 | 19 |
| crim2          | 71  24 | 6 | 71  26 | 3 | 71  27 | 1 | 71  24 | 5 | 83  17 | 0 | 74  23 | 3 | 74  23 | 3 |
| crim1          | 52.8 | 111.3 | 5 | 119.0 | 7 | 103.0 | 70.4 | 101.5 | 101.5 |
| inrs1          | 38  7 | 4 | 51  67 | 8 | 56  3 | 3 | 74  79 | 14 | 95  3 | 1 | 317 327 | 30 | 317 327 | 30 |
| mit_lcsi       | 88  8 | 4 | 75  17 | 8 | 93  2 | 1 | 92  5 | 2 | 94  7 | 4 | 88  4 | 8 | 88  4 | 8 |
| paramax        | 37  11 | 52 | 48  14 | 39 | 46  26 | 20 | 66  7 | 20 | 90  11 | 1 | 58  13 | 29 | 58  13 | 29 |
| srl1           | 74  2.2 | 66.1 | 80.3 | 40.7 | 29.6 | 55.6 | 55.6 |
| srl2           | 69  2 | 9 | 91  19 | 14 | 109  17 | 16 | 144  7 | 16 | 137  7 | 8 | 550 62 | 62 | 413 55 | 54 |
| overall        | 37.6 | 19.0 | 14.5 | 27.6 | 31.4 |

| Totals         | 665 174 | 140 | 883 297 | 184 | 1052 362 | 148 | 1368 240 | 229 | 1389 198 | 85 |
| Foreign        | 36.9 | 19.0 | 14.5 | 27.6 | 31.4 |

Legend:
- AT: ATLAS
- TT: TEL
- L: NLA
- W: Weighted

Table 6: ATIS NL Test Results

| Class (A+D) | Class A | Class D |
|-------------|---------|---------|
| 674 Utt.    | 427 Utt. | 247 Utt. |

| system      | W. Err(%) | W. Err(%) | W. Err(%) |
|-------------|-----------|-----------|-----------|
| att1        | 42.4      | 34.7      | 55.9      |
| bbn1        | 22.0      | 15.7      | 32.8      |
| cmul        | 12.3      | 12.2      | 12.6      |
| crim1       | 71.2      | 40.5      | 124.3     |
| crim2       | 69.4      | 50.1      | 102.8     |
| crim3       | 49.7      | 31.1      | 81.8      |
| inrs1       | 101.5     | 79.9      | 138.9     |
| mit_lcsi    | 18.4      | 18.3      | 18.6      |
| paramax     | 55.6      | 44.0      | 75.7      |
| srl1        | 27.6      | 22.2      | 36.8      |
| srl2        | 23.6      | 14.8      | 38.9      |

Table 7: ATIS NL Results: Class (A+D) by Collection Site
| Class (A+D) Set | Class A | Class D |
|----------------|---------|---------|
| 674 Utt.       | 427 Utt.| 247 Utt.|

| system         | W. Err(%) | W. Err(%) | W. Err(%) | Description                      |
|----------------|-----------|-----------|-----------|----------------------------------|
| att1           | 82.8      | 94.6      | 140.1     | ATTI Nov 92 ATIS SLS Results     |
| bbn1           | 30.6      | 23.7      | 42.5      | BBN1 Nov 92 ATIS SLS Results     |
| cmul           | 21.2      | 19.7      | 23.9      | CMU1 Nov 92 ATIS SLS Results     |
| crml           | 82.3      | 56.9      | 126.3     | CRML CHANEL Nov 92 ATIS SLS Results |
| crml2          | 82.9      | 66.3      | 112.1     | CRML2 CHANEL CD Nov 92 ATIS SLS Results |
| crml3          | 75.2      | 57.1      | 106.5     | CRML3 HEONDON Nov 92 ATIS SLS Results |
| inrsl          | 100.0     | 100.0     | 100.0     | INRSL LATE Nov 92 ATIS SLS Results |
| mit_local      | 29.7      | 30.4      | 28.3      | MIT_LOCAL Nov 92 ATIS SLS Results |
| sr1l           | 37.4      | 31.9      | 47.0      | SRI1 TM Nov 92 ATIS SLS Results   |
| sr12           | 33.2      | 26.5      | 44.9      | SRI2 GEMINI+TM Nov 92 ATIS SLS Results |

Table 8: ATIS SLS Test Results

| Class (A+D) Set | ATTI Nov 92 ATIS SLS Results | BBN Nov 92 ATIS SLS Results | CMU Nov 92 ATIS SLS Results | CRML CHANEL Nov 92 ATIS SLS Results | CRML2 CHANEL CD Nov 92 ATIS SLS Results | CRML3 HEONDON Nov 92 ATIS SLS Results | INRSL LATE Nov 92 ATIS SLS Results | MIT_LOCAL Nov 92 ATIS SLS Results | SRI1 TM Nov 92 ATIS SLS Results | SRI2 GEMINI+TM Nov 92 ATIS SLS Results |
|----------------|----------------------------|----------------------------|-----------------------------|-------------------------------------|---------------------------------------|---------------------------------------|---------------------------------|---------------------------------|---------------------------------|---------------------------------|
|                | 82.8                       | 94.6                       | 140.1                       | ATT                               | BBN                                   | CMU                                   | INRSL                           | MIT                            | SRI1                            | SRI2                            |
|                | 30.6                       | 23.7                       | 42.5                        | BBN1                              | CMU1                                  | CRML1                                 | INRSL1                          | MIT1                           | SRI11                            | SRI2                            |
|                | 21.2                       | 19.7                       | 23.9                        | CMU1                              | CRML1                                 | CRML1                                 | INRSL1                          | MIT1                           | SRI11                            | SRI2                            |
|                | 82.3                       | 56.9                       | 126.3                       | CRML                             | CRML2                                 | CRML2                                 | INRSL1                          | MIT1                           | SRI11                            | SRI2                            |
|                | 75.2                       | 57.1                       | 106.5                       | CMU1                              | CRML2                                 | CRML2                                 | INRSL1                          | MIT1                           | SRI11                            | SRI2                            |
|                | 100.0                      | 100.0                      | 100.0                       | CMU1                              | CRML2                                 | CRML2                                 | INRSL1                          | MIT1                           | SRI11                            | SRI2                            |
|                | 29.7                       | 30.4                       | 28.3                        | CRML1                            | CRML1                                 | CRML2                                 | INRSL1                          | MIT1                           | SRI11                            | SRI2                            |
|                | 37.4                       | 31.9                       | 47.0                        | CRML1                            | CRML2                                 | CRML2                                 | INRSL1                          | MIT1                           | SRI11                            | SRI2                            |
|                | 33.2                       | 26.5                       | 44.9                        | CRML1                            | CRML2                                 | CRML2                                 | INRSL1                          | MIT1                           | SRI11                            | SRI2                            |

Table 9: ATIS SLS Results: Class (A+D) by Collection Site