INTRODUCTION

This paper documents the third in a series of Benchmark Tests for the DARPA Air Travel Information System (ATIS) common task domain. The first results in this series were reported at the June 1990 Speech and Natural Language Workshop [1], and the second at the February 1991 Speech and Natural Language Workshop [2]. The February 1992 Benchmark Tests include: (1) ATIS domain spontaneous speech recognition system tests, (2) ATIS natural language understanding tests, and (3) ATIS spoken language understanding tests.

Since the February 1991 tests, a large ATIS spoken language corpus has been collected, coordinated by a DARPA “Multi-Site ATIS Data Collection Working” (MADCOW) Group. The activities of this group, and NIST’s role in that effort, are documented in another paper in this Proceedings [3].

OCTOBER 1991 “DRY RUN” TESTS

The procedures for test set selection, testing, scoring, adjudication, and reporting for the February 1992 ATIS Benchmark Tests were developed and used for a “dry run” test in October 1991, with unpublished results. A somewhat smaller test set was used at that time, which did not include test data from AT&T. The implementation of the tests was generally regarded as successful within the DARPA MADCOW Group and by the DARPA Spoken Language Program Coordinating Committee.

NEW CONDITIONS FOR THESE TESTS

The structure (and scoring) of these ATIS domain tests differ in several ways from the tests reported at the June 1990 and February 1991 Workshops:

- Following the February 1991 Workshop, minor revisions (e.g., to accommodate connecting flights, clarify terminology, revise headings and restructure tables, improve representation of fare structures, bug fixes, etc) were made to the relational air-travel-information database. The MADCOW data collection effort, and systems developed with this data, made use of this revised relational database (Version 3.3).
- The MADCOW data collection effort provided data from five sites (AT&T, BBN, CMU, MIT/LCS, and SRI), rather than the single ATIS data collection site (TI) used for the June 1990 and February 1991 tests.
- Some (but not all) of the collecting sites provided secondary (Crown PCC-160) microphone data in addition to the primary (close-talking Sennheiser) microphone. The use of the secondary microphone data was encouraged, but not required, for the February 1992 tests.
- The definition of “Class D” queries was broadened to include “Class D1” queries.
- The files indicating the “classification” (i.e., Class A, D or X) for each query were not provided along with the test queries (as they had been in previous tests), so that each site had no extra information regarding the context-dependency or answerability of each query.
- Similarly, “unanswerable” (Class X) queries were not identified when the test material was released. If system developers provided answers for these queries, they were not scored.
- No utterances were to be treated differently on the grounds of the presence of disfluencies such as false starts or restarts. In the February 1991 tests, these utterances were regarded as “Optional”.
- Concern had been expressed at the February 1991 meeting that some sites might have chosen to “over-generate” (by providing verbose) NL and SLS answers rather than provide more succinct answers. It was argued that “correct” answers should have at
least the information in the "ref" files previously used in scoring answers, but no more than in some specified maximal answer. Bob Moore and Eric Jackson, at SRI, proposed and implemented an algorithmic procedure for deriving maximal reference answers ("rf2") from the NLParse-generated SQL files used to generate the .ref files. Bill Fisher at NIST subsequently modified the NIST comparator (used in scoring the NL and SLS results) to implement the new "minimum/maximum" scoring procedure. The Principles of Interpretation document was modified to accommodate these changes.

- Special reports were to be prepared by NIST to partition the tabulations of results according to the originating sites for the test data.
- Following completion of each phase of scoring the results, NIST was to prepare and make available to all participants both detailed and summary reports via anonymous ftp.
- Because there had been a recommendation to report results for all answerable queries in complete subject-scenarios (i.e., the material collected during one subject's working of one scenario), test material was to be provided to the testing sites in complete subject-scenarios. Emphasis was to be placed on analysis of the subset of "answerable" queries (i.e., Class A+D), rather than on the individual classes A and/or D. Further, the weighted error percentage (defined as twice the percentage of incorrect or "false" answers plus the percentage of "No_Answer" responses) was identified as preferable to the single-number "Score" reported at the February 1991 meeting (Score (%) = 100 (%) - [Weighted Error (%)]).

4 TEST MATERIAL SELECTION AND DISTRIBUTION

With the approval of the MADCOW Group, NIST had reserved approximately 20% of the pooled MADCOW data for test purposes. NIST screened this data for the occurrence of truncated utterances, rejected the subject-scenarios that included these phenomena, and determined that there was a sufficient quantity of reserved potential test material to permit release of a test set consisting of approximately 200 utterances from each of the five MADCOW sites contributing data. NIST did not monitor the audio quality of the .wav files nor review the accuracy of the transcriptions, since no criteria for acceptability based on these have been defined, although in retrospect this might have simplified the adjudication process.

The test material, subsequent to deletion of some material during the adjudication process, consisted of 970 non-null (and 1 null) utterances in all classes. The number of distinct scenarios used by all subjects was 42, with a total of 37 subjects ("speakers") completing 122 subject-scenarios. There were 17 male subjects, and 20 were female. Seven of the 122 subject-scenarios used the "Common-1" scenario; however, the test material selected from BBN and CMU did not include any instances of this scenario. The average number of queries per subject-scenario was 8. The MIT subject-scenarios had an average number of 4.6 queries, and SRI and CMU each had an average number of 12.1 queries per subject-scenario. There were 508 lexemes represented in the test material. The average number of words per utterance was about 11.

After NIST selected the test material, it was produced on CD-ROM. The test disc (NIST Speech Disc T3-1.1) was distributed to the testing sites on Jan. 6, 1992.

Concurrent with preparation of the CD-ROMs, NIST staff and the "Annotation Group" at SRI initiated preparation of the annotation files required to implement scoring.

5 TEST PROCEDURE

Following completion of locally administered single-pass-per-system tests, participating sites submitted results for (at least) three ATIS tests: the SPeech REcognition (SPREC), Natural Language (NL) and Spoken Language System (SLS) tests.

The format for data submission via e-mail was specified by NIST and all "official" results were received at NIST by 6:00 AM on Jan. 20, 1992. As in previous ATIS tests, answer hypotheses were to be in the form of lexical SNOR (.lsn) files for the SPREC results and in Common Answer Specification (CAS) format files for the NL and SLS results. Each submission was to be accompanied by a text file for each system providing a system description following a suggested format.

6 TEST SCORING, ADJUDICATION AND REPORTING PROCEDURE

Upon receipt of the test results, NIST implemented preliminary scoring with a reference answer set including .cat, .ref and .rf2 files developed at NIST and SRI for the NL and SLS tests, and the "lexical SNOR" (.lsn) files derived from the detailed (.sro) transcriptions provided by the collecting sites for the SPREC tests. On Jan. 24, 1992, upon completion of the preliminary scoring and
preparation of the required reports, NIST released the preliminary results by anonymous ftp.

A detailed and formal procedure was established at NIST at the MADCOW group's request for handling requests for adjudication.

The participating sites filed a total of 122 requests for adjudication, which were treated by NIST and the SRI Annotation group in a manner similar to that followed for the training data's bug reports. Some of these requests involved more than one utterance, or reported on more than one "bug" in an utterance, so that the number of unique utterances potentially affected by the requests for adjudication was 193, or approximately 19% of the test material.

Of these utterances, the adjudicators determined that 99 (51%) actually required one or more changes. "No Action" decisions were made for the remaining 49%.

NIST was advised by Francis Kubala at BBN during the adjudication period that some of the reference transcriptions used for scoring the SPREC test appeared to be inaccurate. NIST subsequently reviewed all of the transcriptions noted by Kubala and corrected them as deemed appropriate.

In addition to the 99 utterances noted as part of the formal requests for adjudication requiring changes to the annotations, 26 test utterances were identified by the adjudicators as requiring changes.

The final total of 125 utterances (12.9% of the entire test set) for which annotation changes were made includes the following breakdown (by category):

- 42 with software problems related to annotations or scoring (e.g., NLParse, batching, or Comparator bugs),
- 36 for which annotation errors had been made,
- 27 involved problems with the transcriptions developed at the originating sites, and
- 20 involved differences of opinion in applying the Principles of Interpretation or the use of context in interpreting the query.

Following completion of the adjudication process, NIST released a set of "Official" ATIS Benchmark Test results to the community on Feb. 5, 1992.

NIST was subsequently advised by Paramax that corrections to the reference answer set that were to have been made during the adjudication process did not appear to have been made. NIST and SRI determined that this had in fact been the case, and a total of 19 .rf2 files were corrected. The entire set of NL and SLS results were then re-scored, and a "Revised Official" set of results was made available to the community. Analysis of the differences between these two sets of "official" results shows that only 5 of Paramax's NL and 4 of their SLS answers were scored differently.

Paramax also noted, following release of the "Revised Official" results, that 20 of their NL as well as another 20 of their SLS answers were scored as "False" because of known limitations in the NIST official scoring software. NIST had determined that the degree to which Paramax's answers were affected by this known limitation was approximately ten times more severe than for any other site, and declined to alter the scoring software to accommodate Paramax's unusual responses. NIST encouraged Paramax to develop and document "unofficial" results [4] with slightly modified scoring software.

A "handout" was prepared for, and distributed at, the February 1992 Speech and Natural Language Workshop containing the System Descriptions provided by the participants and NIST's summaries of Benchmark Test results.

7 BENCHMARK TEST RESULTS AND DISCUSSION

7.1 ATIS SPEeECH RECOgnition (SPREC) Test Results:

7.1.1 Close-Talking Microphone

Table 1 presents a tabulation of the February 1992 ATIS spontaneous Speech RECOgnition (SPREC) test results.

Results are presented for a number of defined subsets of the utterances, with the utterance classes defined in the annotation process. The set Class A+D+X is the set of all utterances in all classes, consisting of 971 utterances. The set Class A+D includes all answerable utterances, 687 in all. Individual scores for the component subsets Class A, Class D, and Class X are also included. The utterances in Classes D and X tend to have a greater degree of disfluency than those in Class A. This factor may be reflected in the corresponding error rates, since the lowest subset error rates are to be found for Class A utterances, and the highest for Class X.

In the set of answerable queries, Class A+D, the word error ranges from 6.2% to 13.8%, and the "Utterance
error rate" (corresponding approximately to "sentence error rate", but acknowledging the fact that some utterances consist of more than one sentence) range from 34.6% to 60.1%.

The lowest word error rate, in any of the subsets, 5.8%, is noted for the BBN system described in [5] for the subset of Class A utterances.

Table 2 presents a matrix tabulation of ATIS SPREC results for the set of answerable queries, Class A+D. This matrix form of tabulation of results was developed at the MADCOW group's request to shed light on potential variabilities in the data for test set components from differing originating sites. The five columns of the matrix block correspond to the five originating sites for the MADCOW test data. In this case, the six rows of the matrix block correspond to the six sets of SPREC test results sent to NIST. The "Overall Totals" column at the right of the central block presents results corresponding to those cited for the Class A+D subset in Table 1. Note, for example, that the previously cited lowest Class A+D subset word error of 6.2% (for the BBN system) is shown in the second row entry of this column.

The "Overall Totals" row presents results accumulated over all systems for which results were reported to NIST. Note that the Overall (subset) Total Word Error ("W. Error") ranges from a low of 5.9%, for the data originating at MIT/LCS, to 14.6% for the AT&T data subset.

These data suggest that the MIT data subset is less challenging for ATIS SPREC systems than the data from other sites, but the reasons for this are not immediately evident.

Analysis of the transcriptions suggests that the AT&T data subset has a higher incidence of disfluencies than other subsets, partially explaining why it is more challenging than the other data subsets.

For the "Class A+D" data, the lowest subset word error for any SPREC system is 3.2%, again for the BBN SPREC system and for the MIT data subset. Analysis of a similar matrix for the Class A data (not shown) indicates that the lowest subset word error (again for the MIT data subset) is 2.6% for the BBN system, with a corresponding utterance error of 20.7%.

7.1.2 Secondary (Crown PCC-160) Microphone Data

Three ATIS MADCOW sites provided data for both the Sennheiser close-talking microphone and the secondary (Crown PCC-160) microphone: CMU, MIT/LCS, and SRI. Two sites agreed to use the Crown microphone data with SPREC systems, using "robust" recognition algorithms: CMU and SRI. In some cases, results for other algorithms for comparable subsets of the data are available, and these have been excised from larger sets of data provided to NIST by CMU and SRI for the purposes of comparisons.

Table 3 presents a matrix tabulation of the SPREC data for the Class A+D data from CMU, MIT/LCS and SRI for 5 systems (i.e., 3 from CMU and 2 from SRI). The "cmu4" system is the CMU Sphinx II system [6] processing the close-talking microphone data, the "cmu6" system is the CMU codeword-dependent-cepstral-normalization (CDCN) system [7] processing the close-talking data, and the "cmu3" system is the CMU CDCN system processing the Crown microphone data. The "sri3" system (processing the close-talking microphone data) and "sri4" system (processing the Crown microphone data) are versions of the SRI Decipher system incorporating the "RASTA" procedure for high-pass filtering of a log-spectral representation of speech [8].

For the close-talking microphone data subset, the lowest word error rate (7.0%) is for the sri4 system, which may be compared to the cmu4 system (10.4%) and the cmu6 system (13.7%). According to the system description provided by CMU, the two CMU systems differ in the amount of training material, among other factors.

For the secondary microphone data subset, the word error rate for the cmu3 system is 17.8%, and for the sri4 system is 30.4%.

There are indications of substantial variabilities due to originating site for the secondary microphone data, with both the SRI and CMU data secondary microphone data subsets giving rise to higher error rates than for the MIT data subsets.

7.1.3 Statistical Significance: SPREC

As in previous benchmark tests, two statistical significance tests are routinely implemented at NIST in analysis of speech recognition performance assessment tests. The utterance (sentence) error test is an application of McNemar's test, first suggested for use in this community by Gillick [9]. Another test consists of a MAtched-Pairs Sentence-Segment Word Error ("MAPSSWE") significance test, originally devised for use with the Resource Management corpora.

Analysis of the tabulation of the word error test results
for the answerable query subset (Class A+D) shown in Table 4a indicates that for the BBN system [5], the word error rates are significantly different from (lower than) those for the other systems included in these tests. The sentence error McNemar test (Table 4b) indicates a similar result, but in this case, the sentence error rate for the Paramax SPREC system [4] does not differ significantly from the BBN system.

### 7.2 Natural Language (NL) Tests

Table 5 presents a tabulation of the February 1992 ATIS Natural Language (NL) understanding tests results. Results are presented for the set of all “answerable utterances”, Class A+D, and for the individual Class A and Class D subsets. As was the case for the SPREC results, in general the error rates are higher for Class D than for Class A utterances.

For the set of answerable queries, Class A+D, the weighted error ranges from 30.1% to 75.4%. Note that five of the systems have weighted error percentages between 30.1% and 33.9%.

Table 6 presents a matrix tabulation for the NL test results for the set of answerable queries, Class A+D. There were a total of 687 queries in this set. The numbers tabulated for this set in Table 5 appear in the “Overall Totals” column, along with corresponding percentages. The “Overall Totals” row indicates the variability due to the test subsets’ originating site.

Of the 5 data subsets, the lower weighted error percentages in the “Overall Totals” row are to be found for the CMU and MIT data, with the SRI, AT&T, and BBN data giving rise to higher weighted error percentages.

Since the AT&T data was collected using a significantly different collection paradigm – with the subject interfacing with the ATIS system simulation only over a phone line, rather than viewing a screen display of travel information [10] – the fact that the AT&T data subset is more difficult than three other sites is perhaps not surprising.

However, the BBN ATIS data collection effort also differed somewhat from that at other MADCOW sites in that – although information was presented using a screen display – the BBN scenarios “included not only trip planning scenarios, but also problem solving involving more general kinds of database access... This was done to try to elicit a richer range of language usage.[3]” This factor (“richer language usage”) may provide a partial explanation for the high NL error rates noted for the BBN data subset.

For the CMU and MIT [11] systems, there appears to be some indication that the error percentages for “locally-collected” data are lower than for “foreign” data, perhaps because of greater familiarity with the local data-collection scenarios and environment, or use of a variant of the system under test when collecting the MADCOW data from which the test set was selected.

### 7.3 Spoken Language Systems (SLS) Tests

Table 7 presents a tabulation of the February 1992 Spoken Language System understanding test results. As was the case for Table 5 (for the corresponding NL results), results are shown for several classes of the data, but emphasis in this material is placed on the answerable utterances, comprising Class A+D.

For the Class A+D set, the seven SLS systems have weighted error ranging from 43.7% to 90.2%. Note that four systems (from three sites: BBN, MIT and SRI [12]) have weighted error percentages between 43.7% and 52.8%.

Table 8 presents a matrix tabulation for the SLS test results for Class A+D, comparable in structure to that for the NL results of Table 6.

Of the 5 data subsets corresponding to different collection sites, the range in weighted error is from 49.5% (for the MIT data) to 73.1% (for the AT&T data).

### 8 ACKNOWLEDGEMENT

The authors would like to acknowledge the help provided by the entire MADCOW community throughout the ordeal of collecting, annotating, distributing and using the MADCOW corpus for test purposes. A companion paper in this Proceedings provides a detailed acknowledgement, but special credit was earned by Lynette Hirschman as Chair of the MADCOW Group. It is to everyone’s credit that the essential data was collected, annotated, and distributed and that “deadlines” were usually honored!

Special thanks are also due to the group at MIT, particularly Michael Phillips and Christie Clark Winterton, for quick turn- around in producing recordable CD-ROM discs for distribution of the MADCOW training corpus from master tapes produced by Brett Tjaden at NIST.

The Annotation Group at SRI, consisting of Kate Hunicke-Smith, Harry Bratt and Beth Bryson, was invaluable to the NIST effort to implement these tests.
They participated actively and cheerfully in annotation of the test material and the adjudication process, in addition to "training" one of the authors (ND) in the use of the NLParse software and annotation techniques.

Francis Kubala, at BBN, called NIST's attention to some problematic transcriptions for the SPREC tests. NIST reviewed and revised these as appropriate, and in the process noted 3 truncated utterances (in one subject-scenario collected at BBN). While the revised transcriptions were used in NIST's "revised official" scoring, NIST neglected to delete this subject-scenario from the NL and SLS tests, as specified by MADCOW protocols for handling data with truncated utterances. Analysis of performance on this particular subject-scenario indicates that most sites did well, nonetheless.

9 References

1. Pallett, et al., "DARPA ATIS Test Results June 1990", in Proc. Speech and Natural Language Workshop, June 1990, (R. Stern, ed.) Morgan Kaufmann Publishers, Inc. ISBN 1-55860-157-0, pp. 114-121.
2. Pallett, D.S., "Session 2: DARPA Resource Management and ATIS Benchmark Test Poster Session", in Proc. Speech and Natural Language Workshop, February 1991, (P. Price, ed.) Morgan Kaufmann Publishers, Inc. ISBN 1-55860-207-0, pp. 49-58.
3. MADCOW, "Multi-Site Data Collection for a Spoken Language Corpus", in Proc. Speech and Natural Language Workshop, February 1992, (M. Marcus, ed.) Morgan Kaufmann Publishers, Inc.
4. Norton, L.M., Dahl, D.A., and Linebarger, M.C., "Recent Improvements and Benchmark Results for the Paramax ATIS System", in Proc. Speech and Natural Language Workshop, February 1992, (M. Marcus, ed.) Morgan Kaufmann Publishers, Inc.
5. Kubala, F. et al., "BBN BYBLOS and HARC February 1992 ATIS Benchmark Results", in Proc. Speech and Natural Language Workshop, February 1992, (M. Marcus, ed.) Morgan Kaufmann Publishers, Inc.
6. Ward, W. et al., "Speech Understanding in Open Tasks", in Proc. Speech and Natural Language Workshop, February 1992, (M. Marcus, ed.) Morgan Kaufmann Publishers, Inc.
7. Stern, R. M., et al., "Multiple Approaches to Robust speech Recognition" in Proc. Speech and Natural Language Workshop, February 1992, (M. Marcus, ed.) Morgan Kaufmann Publishers, Inc.
8. Murveit, H., Butzberger, J. and Weintraub, M., "Reduced Channel Dependence for Speech Recognition", in Proc. Speech and Natural Language Workshop, February 1992, (M. Marcus, ed.) Morgan Kaufmann Publishers, Inc.
9. Gillick, L. and Cox, S.J., "Some Statistical Issues in the Comparison of speech Recognition Algorithms", Proceedings of ICASSP-89, Glasgow, May 1989, pp.532-535.
10. Pieraccini, R. et al., "Progress Report on the Chronus System: ATIS Benchmark Results", in Proc. Speech and Natural Language Workshop, February 1992, (M. Marcus, ed.) Morgan Kaufmann Publishers, Inc.
11. Zue, V., et al., "The MIT ATIS System: February 1992 Progress Report", in Proc. Speech and Natural Language Workshop, February 1992, (M. Marcus, ed.) Morgan Kaufmann Publishers, Inc.
12. Appelt, D.E. and Jackson, E., "SRI International February 1992 ATIS Benchmark Test Results", in Proc. Speech and Natural Language Workshop, February 1992, (M. Marcus, ed.) Morgan Kaufmann Publishers, Inc.

10 APPENDIX: "OFFICIAL" VS. "UNOFFICIAL" RESULTS

Several sites expressed interest in having results for additional systems included in NIST's "official" summary, although these results typically were not available at the required time for "official" scoring. At least one site took exception to an idiosyncratic property of the "official" comparator's treatment of their system's responses to several queries, and requested permission to present "unofficial" results at the meeting. Another site noted that they had identified a "bug" in their CAS-answer format software, and after it was fixed, they also requested permission to report unofficial results.

It was subsequently decided that the results submitted to NIST by the specified deadline, and uniformly scored at NIST with the "official" comparator and the adjudicated final set of reference answers would comprise the only "official" results, and that locally scored results should be represented as "unofficial", even if scored with the same scoring software and answer set as the "official" results.

It should be noted that since the results are for locally implemented tests, and since NIST's role in the tests is principally one of selecting and distributing the test material, and implementing the scoring software and uniformly tabulating the results of the tests, the results are not to be construed or represented as endorsements of any systems or official findings on the part of NIST, DARPA or the U.S. Government.
### Table 1: ATIS SPREC Test Results

| Subset | Corr | Sub | Del | Ins | Err | U. Err | # Utt. |
|--------|------|-----|-----|-----|-----|--------|-------|
| **Class A+D+X Subset** | | | | | | | |
| att3-adx | 85.6 | 10.5 | 3.9 | 3.1 | 17.5 | 64.6 | 970 |
| bbn3-adx | 92.5 | 5.7 | 1.8 | 1.8 | 9.4 | 40.3 | 971 |
| cmu4-adx | 88.2 | 9.7 | 2.1 | 4.4 | 16.2 | 60.2 | 971 |
| mit4-adx | 84.1 | 11.5 | 4.4 | 2.3 | 18.1 | 59.6 | 971 |
| paramax3-adx | 91.5 | 6.3 | 2.1 | 2.1 | 10.6 | 42.2 | 971 |
| srl3-adx | 91.4 | 6.8 | 1.8 | 2.4 | 11.0 | 48.7 | 971 |
| **Class A Subset** | | | | | | | |
| att3-a | 88.9 | 7.7 | 3.4 | 2.7 | 13.8 | 60.1 | 687 |
| bbn3-a | 95.2 | 3.6 | 1.1 | 1.5 | 6.2 | 34.6 | 687 |
| cmu4-a | 91.9 | 6.5 | 1.6 | 3.7 | 11.8 | 54.4 | 687 |
| mit4-a | 88.3 | 8.7 | 3.1 | 1.9 | 13.6 | 54.1 | 687 |
| paramax3-a | 94.6 | 4.0 | 1.4 | 1.7 | 7.1 | 36.4 | 687 |
| srl3-a | 93.8 | 4.9 | 1.4 | 2.1 | 8.4 | 44.5 | 687 |
| **Class D Subset** | | | | | | | |
| att3-d | 89.0 | 8.7 | 2.3 | 4.1 | 15.2 | 58.9 | 285 |
| bbn3-d | 94.9 | 4.2 | 0.8 | 1.9 | 7.0 | 33.3 | 285 |
| cmu4-d | 90.3 | 8.2 | 1.5 | 4.8 | 14.5 | 54.7 | 285 |
| mit4-d | 86.7 | 10.3 | 3.0 | 2.3 | 15.7 | 53.7 | 285 |
| paramax3-d | 94.1 | 4.7 | 1.1 | 2.2 | 8.1 | 36.1 | 285 |
| srl3-d | 92.5 | 6.4 | 1.1 | 2.8 | 10.3 | 45.3 | 285 |
| **Class X Subset** | | | | | | | |
| att3-x | 77.4 | 17.3 | 5.3 | 3.9 | 26.5 | 75.6 | 283 |
| bbn3-x | 85.5 | 11.0 | 3.5 | 2.7 | 17.2 | 53.9 | 284 |
| cmu4-x | 78.9 | 17.6 | 7.7 | 3.3 | 29.5 | 72.9 | 284 |
| mit4-x | 73.8 | 18.5 | 7.7 | 3.3 | 29.5 | 72.9 | 284 |
| paramax3-x | 83.7 | 12.2 | 4.0 | 3.1 | 19.4 | 56.3 | 284 |
| srl3-x | 85.5 | 11.5 | 3.0 | 2.9 | 17.4 | 58.6 | 284 |

**ATT Feb 92 SPREC Results**

**BBN Feb 92 SPREC Results**

**CMU Feb 92 ATIS Sphinx-II Senn.**

**MIT-LCS Feb 92 SPREC Results**

**Paramax/BBN Feb 92 SPREC Results**

**SRI Feb 92 SPREC Results**
Table 2: ATIS SPREC Results Class A+D by Collection Site

| Originating Site of Test Data | ATT | BBN | CMU | MIT | SRI | Overall | Foreign |
|-------------------------------|-----|-----|-----|-----|-----|---------|---------|
| CMU                           | 9.1 | 11.4| 6.3 | 12.4 | 19.6| 17.8 | 14.5 |
| (101 Utt.)                    |     |     |     |     |     |        |        |
| BBN                           | 14.4| 14.4| 14.4| 6.3 | 12.4| 10.4 | 8.1  |
| (152 Utt.)                    |     |     |     |     |     |        |        |
| CMU                           | 12.4| 12.4| 12.4| 12.4| 12.4| 12.4  | 12.4  |
| (79 Utt.)                     |     |     |     |     |     |        |        |

Table 3: ATIS SPREC Crown and Crown Subset of Sennheiser

Test Results by Collection sites

| Originating Site of Test Data | Overall | Foreign |
|-------------------------------|---------|---------|
| CMU                           |         |         |
| (101 Utt.)                    |         |         |
| BBN                           |         |         |
| (152 Utt.)                    |         |         |
| CMU                           |         |         |
| (79 Utt.)                     |         |         |

22
Table 4: ATIS SPREC Significance Test Comparisons: Class A+D
### Class A-D

| System   | # T | # F | # NA | # Utt | W. Err | Description                  |
|----------|-----|-----|------|-------|--------|------------------------------|
| att1     | 378 | 209 | 100  | 687   | 75.4   | ATT Feb92 ATIS              |
| bbn1     | 527 | 73  | 87   | 687   | 33.9   | BBN Feb92 ATIS              |
| cmu1     | 582 | 102 | 3    | 687   | 30.1   | CMU-Phoenix Feb92 ATIS      |
| cmu8     | 560 | 101 | 26   | 687   | 33.2   | CMU-MINDS-II Feb92 ATIS     |
| mit2     | 551 | 87  | 49   | 687   | 32.5   | MIT Feb92 ATIS              |
| paramax1 | 311 | 122 | 254  | 687   | 72.5   | PARAMAX Feb92 ATIS          |
| sri1     | 533 | 60  | 94   | 687   | 31.1   | SRI Feb92 ATIS              |

### Class A

| System   | # T | # F | # NA | # Utt | W. Err | Description                  |
|----------|-----|-----|------|-------|--------|------------------------------|
| att1-a   | 256 | 56  | 50   | 402   | 60.2   | ATT Feb92 ATIS Class A NL    |
| bbn1-a   | 322 | 26  | 54   | 402   | 26.4   | BBN Feb92 ATIS Class A NL    |
| cmu1-a   | 356 | 48  | 0    | 402   | 22.9   | CMU-Phoenix Feb92 ATIS Class A NL |
| cmu8-a   | 348 | 46  | 10   | 402   | 25.4   | CMU-MINDS-II Feb92 ATIS Class A NL |
| mit2-a   | 342 | 34  | 26   | 402   | 23.4   | MIT Feb92 ATIS Class A NL    |
| paramax1-a | 223 | 50  | 129  | 402   | 57.0   | PARAMAX Feb92 ATIS Class A NL |
| sri1-a   | 335 | 25  | 42   | 402   | 22.9   | SRI Feb92 ATIS Class A NL    |

### Class D

| System   | # T | # F | # NA | # Utt | W. Err | Description                  |
|----------|-----|-----|------|-------|--------|------------------------------|
| att1-d   | 122 | 113 | 50   | 285   | 96.8   | ATT Feb92 ATIS Class D NL    |
| bbn1-d   | 205 | 47  | 33   | 285   | 44.6   | BBN Feb92 ATIS Class D NL    |
| cmu1-d   | 226 | 56  | 3    | 285   | 40.4   | CMU-Phoenix Feb92 ATIS Class D NL |
| cmu8-d   | 214 | 55  | 16   | 285   | 44.2   | CMU-MINDS-II Feb92 ATIS Class D NL |
| mit2-d   | 209 | 53  | 23   | 285   | 45.3   | MIT Feb92 ATIS Class D NL    |
| paramax1-d | 88  | 72  | 125  | 285   | 94.4   | PARAMAX Feb92 ATIS Class D NL |
| sri1-d   | 198 | 35  | 52   | 285   | 42.8   | SRI Feb92 ATIS Class D NL    |

Table 5: Feb 92 ATIS NL Test Results - Using Minimal/Maximal Scoring Criterion
Table 6: ATIS NL Test Results - Using Minimal/Maximal Scoring Criterion
## Class A+D

| System   | # T  | # F  | # NA | # Utt | W. Err | Description                          |
|----------|------|------|------|-------|--------|--------------------------------------|
| att2     | 300  | 233  | 154  | 687   | 50.2   | ATT Feb 92 ATIS SLS                  |
| bbn2     | 493  | 106  | 88   | 687   | 43.7   | BBN Feb 92 ATIS SLS                  |
| cmu2     | 456  | 226  | 3    | 687   | 66.2   | CMU Feb 92 ATIS SLS                  |
| mitl     | 471  | 132  | 84   | 687   | 50.7   | MIT/SRI Feb 92 ATIS SLS              |
| mit3     | 419  | 95   | 173  | 687   | 52.8   | MIT Feb 92 ATIS SLS                  |
| paramax2 | 302  | 148  | 237  | 687   | 77.6   | PARAMAX/BBN Feb 92 ATIS SLS          |
| sr12     | 444  | 69   | 174  | 687   | 45.4   | SRI Feb 92 ATIS SLS                  |

## Class A

| System   | # T  | # F  | # NA | # Utt | W. Err | Description                          |
|----------|------|------|------|-------|--------|--------------------------------------|
| att2-a   | 208  | 118  | 76   | 402   | 77.6   | ATT Feb 92 ATIS SLS Class A           |
| bbn2-a   | 301  | 43   | 58   | 402   | 35.8   | BBN Feb 92 ATIS SLS Class A           |
| cmu2-a   | 298  | 104  | 0    | 402   | 51.7   | CMU Feb 92 ATIS SLS Class A           |
| mitl-a   | 305  | 58   | 39   | 402   | 38.6   | MIT/SRI Feb 92 ATIS SLS Class A       |
| mit3-a   | 288  | 47   | 67   | 402   | 40.0   | MIT Feb 92 ATIS SLS Class A           |
| paramax2-a | 215  | 70   | 117  | 402   | 63.9   | PARAMAX/BBN Feb 92 ATIS SLS Class A   |
| sr12-a   | 303  | 32   | 65   | 402   | 32.1   | SRI Feb 92 ATIS SLS Class A           |

## Class D

| System   | # T  | # F  | # NA | # Utt | W. Err | Description                          |
|----------|------|------|------|-------|--------|--------------------------------------|
| att2-d   | 92   | 115  | 78   | 285   | 108.1  | ATT Feb 92 ATIS SLS Class D           |
| bbn2-d   | 192  | 63   | 30   | 285   | 54.7   | BBN Feb 92 ATIS SLS Class D           |
| cmu2-d   | 160  | 122  | 3    | 285   | 86.7   | CMU Feb 92 ATIS SLS Class D           |
| mitl-d   | 166  | 74   | 45   | 285   | 67.7   | MIT/SRI Feb 92 ATIS SLS Class D       |
| mit3-d   | 131  | 48   | 106  | 285   | 70.9   | MIT Feb 92 ATIS SLS Class D           |
| paramax2-d | 87   | 78   | 120  | 285   | 66.8   | PARAMAX/BBN Feb 92 ATIS SLS Class D   |
| sr12-d   | 139  | 37   | 109  | 285   | 64.2   | SRI Feb 92 ATIS SLS Class D           |

Table 7: ATIS SLS Test Results - Using Minimal/Maximal Scoring Criterion
| Class (A+D) Set | Originating Site of Test Data | Overall | Foreign |
|----------------|-------------------------------|---------|---------|
|                | ATT | BBN | CMU | MIT | SRI |
|                | Totals | Coll. Site | Totals | |
| att2           | 56  | 50  | 28  | 114 | 151 |
|                | 112.3 | 88.8 | 60.3 | 76.3 | 101.5 |
|                 | 50  | 52  | 31  | 150  | 151  |
|                | 122  | 108  | 122  | 122  | 122  |
| bbn2           | 72  | 23  | 19  | 113 | 137 |
|                | 112.3 | 88.8 | 60.3 | 76.3 | 101.5 |
|                 | 50  | 52  | 31  | 150  | 151  |
|                | 122  | 108  | 122  | 122  | 122  |
| cmu2           | 73  | 38  | 3  | 113 | 137 |
|                | 112.3 | 88.8 | 60.3 | 76.3 | 101.5 |
|                 | 50  | 52  | 31  | 150  | 151  |
|                | 122  | 108  | 122  | 122  | 122  |
| cmu1           | 64  | 33  | 3  | 113 | 137 |
|                | 112.3 | 88.8 | 60.3 | 76.3 | 101.5 |
|                 | 50  | 52  | 31  | 150  | 151  |
|                | 122  | 108  | 122  | 122  | 122  |
| sri2           | 55  | 12  | 47  | 112 | 137 |
|                | 112.3 | 88.8 | 60.3 | 76.3 | 101.5 |
|                 | 50  | 52  | 31  | 150  | 151  |
|                | 122  | 108  | 122  | 122  | 122  |

| Overall | 412 | 195 | 191 | 593 | 222 |
| Totals  | 52  | 24  | 24  | 56  | 21  |
|         | 72.8 | 64.9 | 52.9 | 48.8 | 60.5 |

Legend:

- FT: Foreign Test
- FF: Foreign Fail
- FM: Foreign Mistakes
- #T: # of Tests
- #F: # of Failures
- #NA: # of Non-Attacks
- % Weighted Error

Table 8: ATIS SLS Test Results - Using Minimal/Maximal Scoring Criterion