SESSION 2: DARPA RESOURCE MANAGEMENT AND ATIS BENCHMARK TEST POSTER SESSION

David S. Pallett

National Institute of Standards and Technology
Building 225, Room A216
Gaithersburg, MD 20837

I. INTRODUCTION

Following precedents established as early as the March 1987 DARPA Speech Recognition Workshop, previously-unreleased Benchmark Test Material was selected and released to DARPA contractors and others prior to the February 1991 meeting. Results were reported to NIST and scored using "official" scoring software and reference answers and the results were reported to the participants.

All papers in this poster session at the DARPA speech workshop reported results obtained using the Benchmark Test Material. The Workshop Planning Committee suggested a three-part session format consisting of: (1) Introductory remarks, (2) one hour to review and discuss posters, and (3) open discussion.

Section II of this paper presents an overview describing the approaches used by the participants in this session, while Section III summarizes the open discussion. Section IV describes the benchmark test material selection process and benchmark test protocols. Section V presents tabulations of these results, and discussion of these results is included in Section VI.

II. SESSION OVERVIEW

A total of fourteen papers were presented in poster form. Eleven of the papers were presented by DARPA SLS contractors, and three were from non-DARPA sites.

Five papers dealt with speech recognition systems:

(1) The group at Dragon Systems reported speaker-dependent system results for the Resource Management (RM) test set, using the word-pair grammar [1]. Dragon's results were obtained on a 25 Mhz 80486-based PC, with an RM vocabulary modelled using "roughly 30,000 phonemes in context or PICs", and making use of the Dragon rapid match module.

(2) Doug Paul of MIT Lincoln Laboratory reported speech recognition results for both the RM and ATIS SPREC test material [2]. Recent work includes: variations in semiphone modelling, a "very simple improved duration model" responsible for reducing the error rate by about 10%, a new training strategy, and modifications to the recognizer to use back-off bigram language models.

(3) The Spoken Language Group at MIT's Laboratory for Computer Science also reported results for both the RM and ATIS SPREC test material [3]. The MIT SUMMIT system is a "segment-based" speech recognition system, including a front end that incorporates a model of the human peripheral auditory system, a hierarchical segmentation algorithm to identify a network of possible acoustical segments, segmental measurements, and a statistical classifier to produce a phonetic network. The best-scoring word sequence is derived by matching the phonetic network against a pronunciation network. Recent developments have incorporated more complex context-dependency modelling as well as an improved corrective training procedure.

(4) Francis Kubala et al. reported on BBN's BYBLOS results for both the RM and ATIS SPREC test material [4]. The reported RM speaker-independent results include results for a SI model built using only 12 training speakers. BBN's ATIS results include speaker-independent results for two conditions. "The first is a controlled condition using a specific training set and bigram grammar" [similar to that used by Paul [2]]. The second condition makes use of augmented training data (collected at BBN) and a 4-gram class grammar.

(5) A collaborative effort involving Mari Ostendorf and her colleagues at Boston University and others at BBN makes use of a general formalism for integrating two or more speech recognition technologies [5]. "In this formalism, one system uses the N-best search strategy to generate a list of candidate sentences; the list is rescored by other systems; and the different scores are combined to optimize performance." Ostendorf et al. "report on combining the BU system based on stochastic segment models and the BBN system based on hidden Markov models."

Six papers were presented by DARPA contractors describing integrations of speech and natural language processing into ATIS systems.

(1) The Spoken Language Group at MIT's Laboratory for Computer Science presented a status report on the MIT ATIS system [6]. A context-independent version of the SUMMIT system (described in [3]) including a word-pair grammar with perplexity 92 has been incorporated. The back-end has been redesigned, and the parser now produces an intermediate semantic-frame representation "which serves as the focal-point for all back- end operations." Results are reported for both the February '91 ATIS benchmark test set and for a test set collected at MIT.

(2) The Speech and Natural Language Groups at SRI reported results for both the RM and ATIS SPREC speech recognition test sets and for the ATIS NL and SLS tests [7]. The primary emphasis of the SRI presentation was to describe improvements to the SRI DECIPHER speech recognition system, a component in SRI's ATIS system.

Recent "significant" performance improvements are attributed to the addition of tied-mixture HMM modelling. Other approaches discussed include experiments with male-female separation, speaker adaptation, rejection of out-of-vocabulary input, and language modelling (including the use of multi-word lexical units). SRI's "simple serial integration of speech and natural language processing" is said to work well "because the speech recognition system uses a statistical language model to improve recognition performance, and because the natural language processing uses a template matching approach (described elsewhere in this proceedings) that makes it somewhat insensitive to recognition errors".

(3) Wayne Ward presented one of two papers from CMU describing the CMU ATIS System, "PHOENIX" [8]. The speech recognition component consists of a recent vocabulary-independent version of SPHINX, presently without incorporation of out-of- vocabulary models. PHOENIX's "concept of flexible parsing combines frame-based semantics with a semantic phrase grammar," so that the "operation of the parser can be viewed as 'phrase spotting.'" Language modelling included a bigram model for the recognizer and a grammar for the parser.

(4) The second paper from CMU, by Sheryl Young, described the "structure and operation of SOUL (for Semantically-Oriented Understanding of Language)" [9]. SOUL can use semantic and pragmatic knowledge to correct, reject and/or clarify the outputs of the PHOENIX case frame parser in the ATIS domain.

(5) BBN's NL group reported on the BBN DELPHI natural language system and the integration of this system with the BBN BYBLOS system (described in [4]), using an N-best architecture [10]. The BBN authors cite a number of improvements to the DELPHI system that are described in other papers in this Proceedings.

(6) Recent work on the Unisys ATIS Spoken Language System was described by Norton et al. [11]. "Enhancements to the system's semantic processing for handling non-transparent argument structure and enhancements to the system's pragmatic processing of material in answers displayed to the user" are described. In addition to the Unisys system's NL results, results were reported for the case of SLS systems consisting of the Unisys natural language system coupled with two ATIS speech recognition systems: (1) the MIT SUMMIT system (described in [3]) and (2) the MIT Lincoln Labs system (described in [2]). The Unisys system's natural language constraints were also used to select the first-best of N-best speech recognition results (for the SPREC tests) based on syntactic, semantic and pragmatic knowledge.

Three papers were presented by non-DARPA sites.

(1) Douglas O'Shaughnessy described "the initial development of a natural language text processor, as the first step in an INRS [INRS-Telecommunications, University of Quebec] dialogue- by-voice system [12]. A keyword slot-filling approach is used, rather than a "standard parser for English."

(2) In one of two papers from AT&T Bell Laboratories included in this session, Evelyne Tzoukermann described "The Use of a Commercial Natural Language Interface in the ATIS Task" [13]. Tzoukermann relates their "experience in adapting [a commercial natural language interface] to handle domain dependent ATIS queries." The discussion of error analysis notes that, in contrast to the "well-formed" written English for which the commercial product was designed, spontaneous speech contains repetitions, restarts, deletions, interjections and ellipsis, as well as the omission of punctuation marks that "might give the system information".

(3) The second AT&T Bell Laboratories paper, by Pieraccini, Levin and Lee, proposes "a model for a statistical representation of the conceptual structure in a restricted subset of spoken natural language" [14]. The "technique of case decoding" is applied to the Class A sentences in the ATIS domain, with sentences analyzed in terms of 7 general cases: QUERY, OBJECT, ATTRIBUTE, RESTRICTION, Q[UERY] ATTRIBUTE, AND, and DUMMY. Unlike other papers in this session, this paper implements a non-standard test paradigm that prevents explicit comparisons with the results cited for other systems. To address this shortcoming, the authors indicate that they "are developing a module that translates the conceptual representation into an SQL query". Presumably the SQL queries, in conjunction with the ATIS relational database, will permit use of existing DARPA ATIS query-answer performance evaluation procedures.

III. DISCUSSION

Following review of the posters, a number of issues were discussed.

(1) Differences between ATIS Test Sets:

It was noted that there were a number of differences between the June 1990 and February 1991 ATIS test sets, including evidence of greater-than-expected incidence of dysfluencies in the speech and "skewed" or disproportionate representation of some syntactic/semantic phenomena. Doug Paul noted that the test set perplexity for the June 1990 "Class A" test set was 18, in contrast with 22 for the present "Class A" test set, and a perplexity of 45 for the "non-Class A" test material (i.e., all other utterances). Inferences about "progress" or "trends" may thus be complicated by these differences between test sets.

(2) Limited training material:

Also noted was the fact that only a limited amount of fully "canonized" training material—for training acoustic models and for studying such phenomena as dialogue modelling—was available prior to this meeting, in some cases limiting system development. This factor was cited in a number of papers e.g., [2, 4, 8, 10]).

(3) Limitations on the future value of the Resource Management Corpora:

Hy Murveit noted his belief that demonstrable progress in recognizing speaker independent RM1 speech was limited by "how much information we can tease out of [3990] training

utterances". Richard Schwartz took exception to this, citing steady progress in recognizing RM speech.

(4) Properties of ATIS-domain speech:

Richard Schwartz shared some analysis of the ATIS-domain test set speakers. He noted that there was one speaker in the test set with "24 instances of 'uh' in 12 sentences", [which leads to] "a 50% word error rate" for that speaker. On the basis of his analysis, he noted that "people don't know how to talk to a system", and suggested that there ought to be more user/speaker feedback during the data collection process so that the incidence of dysfluencies would be reduced. In response, Hy Murveit noted that if we regard the two worst speakers in the test material as atypical, then "the current word error rate is close to 15%, and with some success in modelling the 'ums' and 'ers', the error rate may be only 10%, or about twice as bad as for RM". Correlation was noted between difficulty in recognizing both the speech and [in understanding] the natural language for the "bad" speakers, so that the suggestion that these speakers may be atypical may be warranted.

Patti Price noted that the [speech recognition] error rates suggest that "ATIS is more difficult, but we don't know why". It may be that ATIS speech is "more casual", but we need to study these issues in more detail, especially as they affect data collection.

(5) Selection of the February ATIS test material:

Victor Zue and Rich Schwartz asked about selection of the February 1991 ATIS test set, asking if there had been screening to select or reject potential test material on the basis of the incidence of dysfluencies noted in the transcriptions. NIST noted that the only such screening was to partition some of the utterances into the "Optional" categories on the basis of evidence of verbal deletions in the "lexical SNOR" transcriptions, since this evidence does not appear in the conventional SNOR transcriptions. For the June test set, there was no such screening, since attention had not been directed to the subset containing verbal deletions.

(6) Use of "Baseline" or "Reference" Conditions:

John Makhoul noted that there "too many uncontrolled variables" (e.g., algorithms, training materials, grammar) to make comparisons of the ATIS speech recognition systems beneficial using the SPREC results. BBN had advocated use of a "baseline" condition and provided SPREC data for both a "baseline" and an "augmented" training condition to permit such comparisons [4]. MIT/LL also made use of this "baseline" condition [2]. Makhoul noted that a similar situation (i.e., "too many uncontrolled variables") applies for the case of the NL results. Hy Murveit noted that SRI's reluctance to "lock into a baseline condition" was based on a reluctance to choose one with the 'wrong operating point'", based on inadequate training. Francis Kubala noted, however, that choosing a "baseline that undershoots" [performance] ought not to be a problem if one wished to "demonstrate clear wins", and that such a baseline could be changed over time. John Makhoul also noted that reporting error rates is in general preferable to reporting "scores".

IV. BENCHMARK TEST MATERIAL AND PROTOCOLS

Benchmark Test Material

One portion of the test material consisted of both "Speaker-Dependent" and "Speaker-Independent" test sets from the Resource Management (RM1) Corpus, for use in tests of speech recognition technology. Each of these test sets consisted of 300 sentence utterances. The most recent tests using the RM1 corpus were conducted prior to the October 1989 Meeting, sixteen months ago. A second portion of the test material consisted of Air Travel Information System (ATIS) domain speech material and related transcriptions. This material was collected at TI in recent months, using the "Wizard" protocol described by Hemphill at the June Meeting [15]. There were a total of 9 speakers in the ATIS test set.

This material was partitioned into four subsets: one subset consisting of an extension of the "Class A" category used at the June meeting (expanded to include "testably ambiguous" queries) and containing 145 queries, a second subset consisting of 38 Class D1 query pairs, and two additional smaller "optional" subsets that included examples of "verbal deletion" and/or "verbal correction" (i.e., Optional Class A and Optional Class D1). The transcriptions used as input to the NL systems and for scoring the ATIS SPREC tests were provided using a recently developed "lexical SNOR" format.

CMU reported benchmark Resource Management results that were not represented in the poster session. These data from CMU are included in the tables of reported results. A paper describing how these results were achieved appears in [17].

Benchmark Test Protocols

In addition to the Resource Management speech recognition tests, for which there is considerable precedent, the ATIS material could be used for three tests: (1) spontaneous ATIS-domain SPeech RECognition component tests (designated as SPREC in this paper), (2) ATIS-domain Natural Language system component tests (designated as NL), and (3) complete ATIS-domain Spoken Language System tests (designated as SLS). During the June meeting, several sites reported results for NL tests, with CMU being the sole site to report complete SLS test results at that meeting [16].

The SPREC test design was outlined by an ad hoc Working Group chaired by Victor Zue, with scoring software adapted for this purpose by NIST. This is the first time that the SPREC test has been implemented.

In computing results tabulated for the NL and SLS tests, the most recent version of the NIST "comparator" was used to compare the hypothesized CAS-format answers against NIST's "canonical" reference answers, as described in a previous paper [16]. Answers are scored as either "True", "False", or (if the No_Answer option has been exercised) "No Answer".

A DARPA SLS Coordinating Committee decision in November, 1990 suggested computation of a "weighted error percentage" on the basis that (on "intuitive grounds") "a false answer is twice as bad as no answer". The weighted error so defined consists of two times the percentage of total queries in the subset that are scored "False" plus the percentage scored

"No Answer". A single-number "score" may be derived by subtracting the weighted error from 100%, providing a single-number score that may range from -100%, for the case of all false answers, to +100% for all true answers.

A "Class D1 test protocol" was developed and used on a trial basis for these tests. Class D1 consists of query pairs for which the second query ("Q2") has been classified as "context dependent", and for which an answerable prior query ("Q1") has been identified as defining the context for Q2. Scoring of Class D1 query pairs was for the answers provided only for Q2, regardless of the answers provided for the context-setting query, Q1.

The Class D1 test protocol had never previously been implemented, and its usage was regarded by many participants more as a "debugging of a test protocol" than as a valid indicator of systems' abilities to handle context-dependent queries. It is also the case that the amount of labelled "Class D1" training material was extremely small and that it was not widely available until shortly before the test —thus limiting system developers' abilities to make adequate use of the training material. Future implementations of the Class D1 test protocol will undoubtedly yield more significant results.

The "optional" test subsets are not discussed extensively in Section VI since these subsets are too small, and their usage too limited, to have significance.

V. BENCHMARK TEST RESULTS

Tables 1 - 4 (included at the end of the text of this paper) present tabulations of results reported to NIST for uniform scoring against the final "official" sets of reference transcriptions and reference answers.

Some of these numbers may differ slightly from those reported at the meeting or in some of the papers in this proceedings, since earlier results reported at the Asilomar meeting were derived with: (1) a slightly larger Class A test set (148 vs. 145 queries), since the classification of 3 utterances, originally included in the Class A subset, was reconsidered, after the meeting, and determined to be "unanswerable" and thus not Class A, and (2) the reference answers for several utterances were corrected and/or modified in response to comments from the participants in these tests. However, these differences are not likely to be statistically significant.

Designation of a set of results as "LATE" signifies that the results were received at NIST some time after midnight on February 6th, 1991. "COB" on that date had been designated as the due date for submission of results. In some cases prior notice had been given to NIST that results would arrive "late", and in a few cases, late results were invited for the sake of completeness and to permit informative comparisons with earlier results.

Resource Management (RM1) Speech Recognition Tests

Table 1 presents a tabulation of speech recognition system results for the (read speech) RM1 test material.

ATIS Spontaneous Speech Recognition Component Tests (SPREC)

Table 2 presents a tabulation of SPREC results for speech recognition systems (or SLS speech recognition components) results for the spontaneous speech in the ATIS domain.

ATIS Natural Language Component Tests (NL)

Table 3 presents a tabulation of natural language system results for the ATIS NL system components (or systems).

In Tables 3 and 4, both the number of queries (and the corresponding percentage of the total number of queries in a given category) are shown for the categories "True" (or correct), "False" (incorrect) and "No Answer". The "Weighted Error" percentage was computed by multiplying the percentage of False answers by 2 and adding the percentage of "No_Answer" responses. The column labelled "Score" was computed by subtracting the Weighted Error (%) from 100%.

ATIS Spoken Language Systems Tests (SLS)

Table 4 presents a tabulation of spoken language system results for complete ATIS systems.

VI. DISCUSSION OF BENCHMARK TEST RESULTS

RM1 Speech Recognition Results (Table 1)

Focusing on the Speaker Independent test set results, with use of the Word Pair grammar, the word error ranges from 9.7% to 3.6%, while the sentence error ranges from 47.3% to 19.3%. Using the NIST implementation of the McNemar test used in previous tests [16], the differences between the sentence error-level results for the system with the lowest reported word and sentence error rates (sys4, the CMU system described in reference [18]) and all other systems in this category are significant for all but sys10 and sys11 (two BBN systems described in reference [4]). The sentence- error-level-performance differences between the CMU system and the two BBN systems are not significant.

There are three sets of results for the BU-BBN collaborative effort described in [5]. The first of these (designated sys7), with a sentence error rate of 27.0%, results from the hybrid BU-BBN system. The second of these (sys12), with a sentence error rate of 27.7%, results from the top answer from the BBN N-best system used for this study. The third (sys13), with a sentence error rate of 47.3%, results from the top answer of the BU context-independent, gender-dependent segment model system.

Lowest overall word and sentence error rates (1.8% and 12.0%, respectively) are reported for the case of the speaker-dependent Word-Pair grammar system results (sys5) reported by Paul, at MIT/LL, described in [2].

In addition to results reported in this session, note that results were reported to NIST for two systems not described in this session. Huang et al. at CMU reported results for an HMM system incorporating a "shared semi-continuous model". That

system is described in a paper to be presented at ICASSP-91 [17]. Gauvain and Lee at AT&T Bell Laboratories reported results for an investigation "into the use of Bayesian learning of the parameters of a Gaussian mixture density", and this study is reported in another paper in this proceedings [18].

ATIS SPREC Results (Table 2)

Focusing on the word error for the 145 utterances in the Class A test set, the range is from 46.1% to 15.7%, while the sentence error ranges from 91.0% to 52.4%.

The McNemar sentence-error-level significance test (not shown) indicates that the system with the lowest reported word and sentence error rate for the Class A utterances (sys24-a, the Unisys implementation of syntactic, semantic and pragmatic constraints in selecting the first candidate from an N-best listing provided by BBN, described in [11]) has an error rate that is significantly less than all but two other systems, (sys18-a, the BBN "augmented training" system, and sys06-a, the SRI system). Performance differences (at the sentence error level) between these three systems, however, are not significant.

Comparison of the results for the BBN "baseline" and "augmented" training condition (sys18 and sys19) gives some indication of the benefits of additional (in this case, domain-specific) training and a more powerful 4-gram statistical class grammar. The McNemar test indicates that the difference in performance between sys18-a and 19-a is significant.

Comparisons of results for similar systems for the two larger test subsets (i.e., Class A results vs. Class D1 results) suggest that the Class D1 material is somewhat more difficult to recognize (i.e., the error rates are higher). An interesting hypothesis that may account, in part, for this phenomenon is offered by Norton et al.: "...this higher error rate in context dependent spontaneous utterances may be due in part to the presence of prosodic phenomena common in dialogue such as destressing 'old' information" [11].

Typical SPREC error rates are higher still for the two "optional" test subsets. This ought not to be surprising in view of the fact that these utterance subsets are, by definition and selection, more dysfluent (i.e., contain verbal deletions).

Not shown in Table 2, but indicated by other analyses, is high inter-subject variability for the SPREC tests as well as for the NL and SLS tests.

ATIS NL Results (Table 3)

For the Class A subset, results are tabulated for eight NL systems at 5 DARPA contractors' sites, and at AT&T Bell Laboratories and at INRS-Telecom. For the DARPA contractor's systems, the weighted error ranges from 51.7% to 31.0%.

The two sets of CMU results include data for the PHOENIX system described in [8] (sys01), and for the PHOENIX system integrated with the SOUL module described by Young in [9] (sys02).

For the Class A test material, the lowest weighted error figures (31.0%) are found for both the SRI system described by Murveit et al. in [7] (sys13-a), and for the CMU PHOENIX + SOUL system of [9] (sys02-a).

For the Class D1 and Optional Class D1 subsets, the weighted error percentages are substantially higher than for the Class A results. For the Class D1 test material, the lowest weighted error figures (36.8%) are found for the Unisys system described by Norton et al. in [11] (sys09-d).

Note that two sets of results are reported for the Class D1 material for BBN (denoted sys15-d and sys23-d). Subsequent to submission of the initial results for sys15-d, BBN's representatives notified NIST that "...there was a small bug in the component that translates the result of the understanding (i.e., the output of the discourse component) into SQL... [and that since] the bug in our system... was NOT in the UNDERSTANDING or the DISCOURSE component but between the output of those components and the SQL backend and ... [since] one small quick fix in the backend corrected the problem, we concluded that it is reasonable to send you new answers for our Class D test" [19]. The data designated as sys23-d is derived from these "new answers".

ATIS SLS Results (Table 4)

For the Class A subset, results are tabulated for 7 SLS systems at 5 DARPA contractors' sites. Non-DARPA contractors declined to participate in the SLS tests. The weighted error ranges considerably, from 90.3% to 41.4%, with the best (lowest weighted error) results for the SRI system described in [7] and in other SRI papers in this proceedings.

The low SRI SLS weighted error rate (41.4%) appears to be a consequence of both a well-performing ATIS speech recognition component and a well-performing natural language component (i.e., a SPREC test word error rate of 18.0% and an NL weighted error rate of 31.0%).

Not surprisingly, weighted error figures for complete SLS systems are higher than for corresponding NL components (processing the lexical SNOR formatted versions of the same utterances). The relative increase in weighted error rate appears to correspond to the relative performance of the speech recognition component.

By comparing comparable data from Tables 3 and 4, note that for the SRI system the weighted error rate for the Class A subset increases from 31.0% (for the NL component) to 41.4% (for the complete SLS system.

Two SLS systems made use of BBN's BYBLOS ATIS SPREC data: the BBN HARC system (sys16-a) and the Unisys-BBN SPREC system (sys22-a). Comparing the increases of weighted error rates for NL vs. SLS systems, one can note an approximate increase in weighted error rate of only 8 or 9 percentage points for these systems (i.e., from 49.0% for the BBN DELPHI NL system to 57.2% for the BBN HARC SLS system, and from 51.7% for the Unisys NL system to 60.7% for the Unisys-((BBN SPREC)) SLS system). This relatively small increase in error rate is probably attributable to the BBN "augmented training" (sys18-a) SPREC test word error rate of (only) 16.1%, which is not significantly different from SRI's SPREC test results of 18.0%.

In contrast, a substantially larger increase in error rate can be noted for the CMU systems (i.e., 35.9% and 31.0% for the two CMU NL systems vs. 65.5% for the SLS system), probably due to performance of the CMU SPREC system with error rates that are significantly higher than for the SRI SPREC system.

Unisys reported results for three system configurations: using speech recognition results provided by the MIT/LCS ATIS SPREC system (designated sys10-a), by the MIT/LL ATIS SPREC system (sys 11-a), and by the BBN BYBLOS/ATIS system (sys22-a). In this case, better performance on the SLS test (i.e., lower weighted error) correlates with better performance on the SPREC results, as would be expected.

As was also the case for the NL results, the weighted error results for the Class D1 test subset are substantially higher than for the Class A results.

VII. ACKNOWLEDGEMENT

Too many individuals have served as points-of-contact at the research sites involved in these benchmark tests to be individually thanked, but their efforts and patience in seeing that information and data are made available are greatly appreciated. My colleagues at NIST deserve special thanks for their efforts and efficiency in making these tests possible and in tabulation of the results. In particular, Bill Fisher has had a key role, both as Chairman of the DARPA SLS Performance Evaluation Working Group and as the individual responsible for ATIS test material selection and in reviewing the "canonical" auxiliary files. Jon Fiscus and John Garofolo, also at NIST, have been responsible for implementation of scoring software and for preparation of corpora on CD-ROM.

VIII. REFERENCES

1. Baker, J., et al., "Dragon Systems Resource Management Benchmark Test Results—February 1991" (in this Proceedings).

2. Paul, D. B. "New Results with the Lincoln Tied-Mixture HMM CSR System" (in this Proceedings).

3. Phillips, M., Glass, J. and Zue, V., "Modelling Context Dependency in Acoustic-Phonetic and Lexical Representations" (in this Proceedings).

4. Kubala, F. et al., "BYBLOS Speech Recognition Benchmark Results" (in this Proceedings).

5. Ostendorf, M. et al., "Integration of Diverse Recognition Methodologies Through Reevaluation of N-Best Sentence Hypotheses" (in this Proceedings).

6. Seneff, S. et al., "Development and Preliminary Evaluation of the MIT ATIS System" (in this Proceedings).

7. Murveit, H. et al., "SRI's Speech and Natural Language Evaluation" (in this Proceedings).

8. Ward, W., "Current Status of the CMU ATIS System" (in this Proceedings).

9. Young, S., "Using Semantics to Correct Parser Output for ATIS Utterances" (in this Proceedings).

10. Austin, S. et al., "BBN HARC and Delphi Results on the ATIS Benchmarks—February 1991" (in this Proceedings).

11. Norton, L. et al., "Augmented Role filling Capabilities for Semantic Interpretation of Spoken Language" (in this Proceedings).

12. O'Shaughnessy, D., "A Textual Processor to Handle ATIS Queries" (in this Proceedings).

13. Tzoukermann, E., "The Use of a Commercial Natural Language Interface in the ATIS Task" (in this Proceedings).

14. Pieraccini, R., Levin, E. and Lee, C.H., "Stochastic Representation of Conceptual Structure in the ATIS Task" (in this Proceedings).

15. Hemphill, C.T., Godfrey, J.J., and Doddington, G.R., "The ATIS Spoken Language System Pilot Corpus" in Proceedings of the DARPA] Speech and Natural Language Workshop" June 1990, pp. 96 -101.

16. Pallett, D.S., et al. "DARPA ATIS Test Results: June 1990" in Proceedings of the DARPA] Speech and Natural Language Workshop" June 1990, pp. 114 - 121.

17. Huang, X. et al., "Improved Acoustic Modelling for the SPHINX Speech Recognition System", (to be presented at ICASSP-91).

18. Gauvain, J. and Lee, C.H., "Bayesian Learning of Gaussian Mixture Densities for Hidden Markov Models" (in this Proceedings).

19. ARPANET communication from M. Bates and R. Ingria (BBN) to Dave Pallett (NIST), February 13, 1991.

FEB91 RM1 SPEECH RECOGNITION TEST

SPEAKER-INDEPENDENT WITHOUT GRAMMAR

NIST-ID	Corr	Sub	Del	Ins	Err	S.Err	Arr. Date	Description
sys1	84.3	12.5	3.2	1.8	17.6	66.0	Jan-31	SRI Spkr-Indep no grammar
sys4	86.2	11.5	2.3	3.2	17.0	66.7	Feb-5	CMU Spkr-Indep no grammar
sys6	83.2	14.2	2.7	2.9	19.7	71.7	Feb-11	MIT-LL Spkr-Indep no grammar-LATE
sys8	81.9	14.8	3.3	2.5	20.7	74.7	Feb-6	AT&T Spkr-Indep no grammar
sys9	82.2	14.4	3.4	2.0	19.8	70.7	Feb-7	AT&T-R Spkr-Indep no grammar-LATE
sys10	83.3	13.6	3.1	2.1	18.8	69.3	Feb-7	BBN Spkr-Indep (109) no grammar-LATE

SPEAKER-INDEPENDENT WORD-PAIR GRAMMAR

NIST-ID	Corr	Sub	Del	Ins	Err	S.Err	Arr. Date	Description
sys1	95.9	3.0	1.0	0.8	4.8	26.0	Jan-31	SRI Spkr-Indep Word-Pair
sys2	93.3	6.0	0.7	1.2	8.0	33.7	Feb-4	MIT Spkr-Indep Word-Pair
sys4	96.8	2.5	0.8	0.4	3.6	19.3	Feb-5	CMU Spkr-Indep Word-Pair
sys6	96.2	2.8	1.0	0.6	4.4	23.3	Feb-6	MIT-LL Spkr-Indep Word-Pair
sys7	95.7	3.3	1.0	1.2	5.6	27.0	Feb-6	BU-BBN Spkr-Indep Word-Pair
sys8	95.5	3.5	1.0	0.7	5.2	28.0	Feb-6	AT&T Spkr-Indep Word-Pair
sys10	96.7	2.3	0.9	0.5	3.8	21.0	Feb-7	BBN Spkr-Indep (109) Word-Pair-LATE
sys11	96.7	2.8	0.6	0.5	3.8	23.0	Feb-7	BBN Spkr-Indep (12) Word-Pair-LATE
sys12	95.7	3.3	1.0	1.1	5.4	27.7	Feb-8	BU-BBN (W/O BU SSM) Spkr-Indep Word-Pair-LATE
sys13	93.0	5.3	1.8	2.6	9.7	47.3	Feb-12	BU Segment Model Spkr-Indep Word-Pair-LATE
sys14	96.1	3.0	0.8	0.7	4.5	25.7	Feb-28	AT&T Sex-Modelled Spkr-Indep Word-Pair-LATE

SPEAKER-DEPENDENT WITHOUT GRAMMAR

NIST-ID	Corr	Sub	Del	Ins	Err	S.Err	Arr. Date	Description
sys5	92.5	5.8	1.7	1.3	8.7	44.0	Feb-6	MIT-LL Spkr-Dep no grammar

SPEAKER-DEPENDENT WORD-PAIR GRAMMAR

NIST-ID	Corr	Sub	Del	Ins	Err	S.Err	Arr. Date	Description
sys3	94.1	4.5	1.4	1.5	7.5	34.3	Feb-5	Dragon Spkr-Dep Word-Pair
sys5	98.3	1.0	0.7	0.1	1.8	12.0	Feb-6	MIT-LL Spkr-Dep Word-Pair

TABLE 1.

Key to Table 1: The following key is provided as an aid in cross-referencing the NIST-ID numbers to the sites submitting results and to descriptions of the systems in the references cited in this paper.

KEY: RM1 SPEECH RECOGNITION TEST REFERENCES

NIST-ID	Site	Reference	NIST-ID	Site	Reference
sys1	SRI	[7]	sys8	AT&T	[18]
sys2	MIT-LCS	[3]	sys9	AT&T	[18]
sys3	Dragon	[1]	sys10	BBN	[4]
sys4	CMU	[17]	sys11	BBN	[4]
sys5	MIT-LL	[2]	sys12	BU-BBN	[5]
sys6	AT&T	[2]	sys13	BU-BBN	[5]
sys7	BU-BBN	[5]	sys14	AT&T	[18]

Feb91 ATIS SPREC Test

Class A

NIST-ID	Corr	Sub	Del	Ins	Err	S.Err	Arr. Date	Description
sys04-a	73.6	19.5	7.0	2.6	29.0	79.3	Feb-6	CMU Class-A SPREC
sys05-a	79.8	16.2	4.0	5.9	26.1	88.3	Feb-6	MIT-LL Class-A SPREC
sys06-a	86.4	10.5	3.1	4.4	18.0	60.0	Feb-6	SRI Class-A SPREC
sys08-a	65.2	28.1	6.8	8.8	43.6	86.2	Feb-6	MIT-LCS Class-A SPREC
sys14-a	63.3	30.4	6.3	9.4	46.1	91.0	Feb-6	Unisys/MIT-LCS Class-A SPREC
sys18-a	87.6	9.9	2.6	3.7	16.1	54.5	Feb-11	BBN Class-A SPREC ("augmented")-LATE
sys19-a	82.5	14.5	3.0	5.3	22.8	67.6	Feb-11	BBN Class-A SPREC ("baseline")-LATE
sys24-a	88.4	9.4	2.2	4.1	15.7	52.4	Feb-13	Unisys/BBN Class-A Sprec-LATE

Class D1

NIST-ID	Corr	Sub	Del	Ins	Err	S.Err	Arr. Date	Description
sys04-d	73.7	17.6	8.6	0.7	26.9	77.6	Feb-6	CMU Class-D1 SPREC
sys05-d	70.7	22.2	7.1	3.9	33.2	91.4	Feb-11	MIT-LL Class-D1 SPREC-LATE
sys06-d	78.6	16.4	4.9	1.2	22.5	67.2	Feb-6	SRI Class-D1 SPREC
sys14-d	52.0	34.1	13.9	6.6	54.6	91.4	Feb-6	Unisys/MIT-LCS Class-D1 SPREC
sys18-d	81.5	13.4	5.1	2.0	20.5	58.6	Feb-11	BBN Class-D1 SPREC ("augmented")-LATE
sys19-d	77.1	17.3	5.6	3.4	26.3	74.1	Feb-11	BBN Class-D1 SPREC ("baseline")-LATE
sys24-d	82.4	13.2	4.4	2.4	20.0	58.6	Feb-13	Unisys/BBN Class-D1 Sprec-LATE

Optional Class A

NIST-ID	Corr	Sub	Del	Ins	Err	S.Err	Arr. Date	Description
sys04-ao	61.7	29.6	8.7	5.3	43.7	81.8	Feb-6	CMU Optional Class-A SPREC
sys05-ao	74.3	22.8	2.9	13.1	38.8	100.0	Feb-11	MIT-LL Optional Class-A SPREC-LATE
sys06-ao	76.7	18.9	4.4	6.8	30.1	90.9	Feb-6	SRI Optional Class-A SPREC
sys14-ao	51.5	42.2	6.3	14.6	63.1	100.0	Feb-6	Unisys/MIT-LCS Optional Class-A SPREC
sys18-ao	74.8	23.8	1.5	11.7	36.9	90.9	Feb-11	BBN Optional Class-A SPREC ("augmented")-LATE
sys19-ao	74.8	23.3	1.9	16.0	41.3	100.0	Feb-11	BBN Optional Class-A SPREC ("baseline")-LATE
sys24-ao	75.2	23.3	1.5	12.6	37.4	90.9	Feb-13	Unisys/BBN Optional Class-A Sprec-LATE

Optional Class D1

NIST-ID	Corr	Sub	Del	Ins	Err	S.Err	Arr. Date	Description
sys04-do	73.7	22.8	3.5	7.0	33.3	75.0	Feb-6	CMU Optional Class-D1 SPREC
sys05-do	80.7	15.8	3.5	21.1	40.4	100.0	Feb-11	MIT-LL Optional Class-D1 SPREC-LATE
sys06-do	87.7	10.5	1.8	22.8	35.1	100.0	Feb-6	SRI Optional Class-D SPREC
sys14-do	59.6	40.4	0.0	15.8	56.1	100.0	Feb-6	Unisys/MIT-LCS Optional Class-D1 SPREC
sys18-do	82.5	14.0	3.5	21.1	38.6	75.0	Feb-11	BBN Optional Class-D1 SPREC ("augmented")-LATE
sys19-do	82.5	15.8	1.8	14.0	31.6	100.0	Feb-11	BBN Optional Class-D1 SPREC ("baseline")-LATE
sys24-do	84.2	12.3	3.5	24.6	40.4	75.0	Feb-13	Unisys/BBN Optional Class-D1 Sprec-LATE

TABLE 2.

FEB91 ATIS NL TEST

Class A

NIST-ID	True		False		No Ans.		W. Err	Score	Arr. Date	Description
sys01-a	117	(80.6%)	24	(16.5%)	4	(2.7%)	35.9	64.1	Feb-6	CMU Class-A NL
sys02-a	117	(80.6%)	17	(11.7%)	11	(7.5%)	31.0	69.0	Feb-6	CMU Class-A NL with knowledge-based module
sys07-a	82	(56.5%)	2	(1.3%)	61	(42.0%)	44.8	55.2	Feb-6	MIT-LCS Class-A NL
sys09-a	84	(57.9%)	14	(9.6%)	47	(32.4%)	51.7	48.3	Feb-6	Unisys Class-A NL
sys12-a	69	(47.5%)	60	(41.3%)	16	(11.0%)	*	*	Feb-7	AT&T Class-A NL-LATE
sys13-a	109	(75.1%)	9	(6.2%)	27	(18.6%)	31.0	69.0	Feb-6	SRI Class-A NL
sys15-a	85	(58.6%)	11	(7.5%)	49	(33.7%)	49.0	51.0	Feb-7	BBN Class-A NL (DELPHI)-LATE
sys17-a	56	(38.6%)	89	(61.3%)	0	(0%)	*	*	Feb-7	INRS Class-A NL-LATE

Class D1

NIST-ID	True		False		No Ans.		W. Err	Score	Arr. Date	Description
sys01-d	25	(65.7%)	11	(28.9%)	2	(5.2%)	63.2	36.8	Feb-6	CMU Class-D1 NL
sys02-d	25	(65.7%)	6	(15.7%)	7	(18.4%)	50.0	50.0	Feb-6	CMU Class-D1 NL with knowledge-based module
sys07-d	18	(47.3%)	2	(5.2%)	18	(47.3%)	57.9	42.1	Feb-6	MIT-LCS Class-D1 NL
sys09-d	24	(63.1%)	0	(0%)	14	(36.8%)	36.8	63.2	Feb-6	Unisys Class-D1 NL
sys12-d	17	(44.7%)	18	(47.3%)	3	(7.8%)	*	*	Feb-7	AT&T Class-D1 NL-LATE
sys13-d	22	(57.8%)	3	(7.8%)	13	(34.2%)	50.0	50.0	Feb-6	SRI Class-D1 NL
sys15-d	10	(26.3%)	3	(7.8%)	25	(65.7%)	81.6	18.4	Feb-7	BBN Class-D1 NL (DELPHI)/BUG-LATE
sys23-d	26	(68.4%)	3	(7.8%)	9	(23.6%)	39.5	60.5	Feb-13	BBN Class-D1 NL (DELPHI)/DEBUGGED-LATE

Optional Class A

NIST-ID	True		False		No Ans.		W. Err	Score	Arr. Date	Description
sys09-ao	1	(9.0%)	0	(0%)	10	(90.9%)	90.9	9.1	Feb-6	Unisys Optional Class-A NL
sys12-ao	2	(18.1%)	8	(72.7%)	1	(9.0%)	*	*	Feb-7	AT&T Optional Class-A NL-LATE
sys13-ao	3	(27.2%)	1	(9.0%)	7	(63.6%)	81.8	18.2	Feb-6	SRI Optional Class-A NL
sys17-ao	3	(27.2%)	8	(72.7%)	0	(0%)	*	*	Feb-7	INRS Optional Class-A NL-LATE

Optional Class D1

NIST-ID	True		False		No Ans.		W. Err	Score	Arr. Date	Description
sys09-do	0	(0%)	1	(50.0%)	1	(50.0%)	150.0	-50.0	Feb-6	Unisys Optional Class-D1 NL
sys12-do	0	(0%)	2	(100.0%)	0	(0%)	*	*	Feb-7	AT&T Optional Class-D1 NL-LATE
sys13-do	0	(0%)	2	(100.0%)	0	(0%)	200.0	-100.0	Feb-6	SRI Optional Class-D1 NL

* Non-DARPA contractors, unpublished by mutual agreement

TABLE 3.

FEB91 ATIS SLS TEST

Class A

NIST-ID	True		False		No Ans.		W. Err	Score	Arr. Date	Description
sys03-a	89	(61.3%)	39	(26.8%)	17	(11.7%)	65.5	34.5	Feb-6	CMU Class-A SLS
sys10-a	29	(20.0%)	15	(10.3%)	101	(69.6%)	90.3	9.7	Feb-6	Unisys Class-A SLS (MIT-LCS SPREC)
sys11-a	32	(22.0%)	5	(3.4%)	108	(74.4%)	81.4	18.6	Feb-6	Unisys Class-A SLS (MIT-LL SPREC)
sys16-a	84	(57.9%)	22	(15.1%)	39	(26.8%)	57.2	42.8	Feb-11	BBN Class-A SLS (HARC)-LATE
sys20-a	46	(31.7%)	19	(13.1%)	80	(55.1%)	81.4	18.6	Feb-6	MIT-LCS Class-A SLS
sys21-a	96	(66.2%)	11	(7.5%)	38	(26.2%)	41.4	58.6	Feb-6	SRI Class-A SLS
sys22-a	77	(53.1%)	20	(13.7%)	48	(33.1%)	60.7	39.3	Feb-13	Unisys Class-A SLS (BBN SPREC)-LATE

Class D1

NIST-ID	True		False		No Ans.		W. Err	Score	Arr. Date	Description
sys03-d	16	(42.1%)	20	(52.6%)	2	(5.2%)	110.5	-10.5	Feb-6	CMU Class-D1 SLS
sys21-d	15	(39.4%)	11	(28.9%)	12	(31.5%)	89.5	10.5	Feb-6	SRI Class-D1 SLS
sys26-d	7	(18.4%)	3	(7.8%)	28	(73.6%)	89.5	10.5	Feb-16	BBN Class-D1 SLS (HARC)-LATE

Optional Class A

NIST-ID	True		False		No Ans.		W. Err	Score	Arr. Date	Description
sys10-ao	2	(18.1%)	0	(0%)	9	(81.8%)	81.8	18.2	Feb-6	Unisys Optional Class-A SLS (MIT-LCS SPREC)
sys21-ao	3	(27.2%)	0	(0%)	8	(72.7%)	72.7	27.3	Feb-6	SRI Optional Class-A SLS
sys22-ao	2	(18.1%)	0	(0%)	9	(81.8%)	81.8	18.2	Feb-13	Unisys Optional Class-A SLS (BBN SPREC)-LATE

Optional Class D1

NIST-ID	True		False		No Ans.		W. Err	Score	Arr. Date	Description
sys21-do	0	(0%)	1	(50.0%)	1	(50.0%)	150.0	-50.0	Feb-6	SRI Optional Class-D1 SLS

TABLE 4.

Key to Tables 2, 3 and 4: The following key is provided as an aid in cross-referencing the NIST-ID numbers to the sites submitting ATIS results and to descriptions of the systems in the references cited in this paper. Note: key for these tables differs from that for the RM1 results of Table 1.

KEY: ATIS SPREC, NL, AND SLS TEST REFERENCES

NIST-ID	Site	Reference	NIST-ID	Site	Reference
sys01	CMU	[8]	sys13	SRI	[7]
sys02	CMU	[9]	sys14	Unisys	[11]
sys03	CMU	[8]	sys15	BBN	[10]
sys04	CMU	[8]	sys16	BBN	[10]
sys05	MIT-LL	[2]	sys17	INRS	[12]
sys06	SRI	[7]	sys18	BBN	[4]
sys07	MIT-LCS	[6]	sys19	BBN	[4]
sys08	MIT-LCS	[3]	sys20	MIT-LCS	[6]
sys09	Unisys	[11]	sys21	SRI	[7]
sys10	Unisys	[11]	sys22	Unisys	[11]
sys11	Unisys	[11]	sys23	BBN	10]
sys12	AT&T	[13]	sys24	Unisys	[11]