

DARPA ATIS Test Results

June 1990

D. S. Pallett, W. M. Fisher, J. G. Fiscus, and J. S. Garofolo

Room A 216 Technology Building
National Institute of Standards and Technology (NIST)
Gaithersburg, MD 20899

Introduction

The first Spoken Language System tests to be conducted in the DARPA Air Travel Information System (ATIS) domain took place during the period June 15 - 20, 1989. This paper presents a brief description of the test protocol, comparator software used for scoring results at NIST, test material selection process, and preliminary tabulation of the scored results for seven SLS systems from five sites: BBN, CMU, MIT/LCS, SRI and Unisys. One system, designated cmu-spi(r) in this paper, made use of digitized speech as input (.wav files), and generated CAS-format answers. Other systems made use of SNOR transcriptions (.snr files) as input.

Test Protocol

The test protocol for these tests was modelled after precedents established over the past several years in the DARPA Resource Management speech recognition benchmark tests. On June 11, 1990, participating sites were notified of the availability of SNOR-format transcriptions for a designated set of 93 "Class A" test utterances. Data were available using an FTP protocol that had been used earlier for access to system training material. Copies of the speech waveform files were distributed to three sites using Exabyte tapes. Responses were provided (in most cases) to NIST on June 15. In return, NIST

provided a key to an encrypted version of the complete ATIS session data from which the test material was selected, and sites were free to access that data in scoring their system results locally. A preliminary summary of the test results was distributed by NIST to participants on June 18th.

Availability of the ATIS Data

Access to all of the ATIS data released by NIST (except for the speech waveform data) has been available via anonymous FTP. The speech waveform data has been made available to three sites to date on 8mm Exabyte tape: AT&T, CMU and SRI. Production of the entire Pilot ATIS Corpus is planned for release by NIST on CD-ROM media after completion of the corpus collection effort at TI.

Comparator Software

Both REF (reference) and HYP (hypothesized) answers were to be written in a CAS (Common Answer Specification) format that was a slight adaptation of the CAS originally developed by BBN, and which had been agreed on by the CAS/Comparator Task Group. Two programs were available to aid in the evaluation by automatically comparing matching REF and HYP answers: one in LISP, contributed by BBN, and one in C, developed at NIST. Final responsibility for decisions on

whether or not an answer is correct rested with human judges at NIST.

When the comparator programs ran, there were a few disagreements in two areas: (1) some answers were scored "correct" by the more forgiving NIST code even though the REF and HYP answers disagreed in the use of quotation marks delimiting an answer; and (2) some HYP answers consisting of tables of numeric codes were incorrectly counted as matching the REF answers by the NIST code.

The first area is a trivial matter of formats and how forgiving a program should be. Our judgement in these cases was that the content of the answer was correct.

The second area raises some interesting logical questions. In a typical case, the required column in the table was a code that looked like a number, such as `flight_code`; because integers and floating point numbers were to be treated as the same type, the tolerance for floating point comparison was used in deciding equality. Because the key fields of the extra erroneous rows were "close enough" to the correct ones, they were ignored. An ad hoc code change was subsequently made in the NIST code so that the tolerance was used only in equality tests when at least one of the numbers was floating point. But we think that in principle there is nothing wrong about using a tolerance in comparing two integers; the real wrongness is treating a pointer (or name) as a number. One principled way to clear this up would be to consistently use enclosing quotation marks to indicate tokens that are not to be treated numerically.

We had to increase considerably the space allotted to input buffers in the NIST C software, since one answer that was submitted took more than 175 K bytes.

As a result of seeing some particular answers, one more change was made in the NIST Comparator code to make it more forgiving: leading and trailing whitespace in a string is

now ignored. This made several answers from one of the sites count as correct, in agreement with our judgement that they had the right content.

Several examples came up in the test answers to illustrate the trouble with looking for matches of only values, without constraining the values to be of the same variable, in conjunction with allowing extra values in a tuple. For instance, query `bd00c1s`, "WHAT IS CLASS Y", has the REF answer `(("Y"))`, and one of the HYP answers supplied is:

```
(("Y" "Y" "COACH" "NO" "YES" "NO" "NO"
 "NONE" "1234567"))
```

Our CAS specification counts this as a correct match, although it is indeterminate which of the tuple's two "Y" fields was matched. A HYP answer with "Y" value in any field would count as equally correct. In tabular answers intended for human consumption, this problem is solved by supplying column headings. It would be easy to incorporate a similar system into our computerized scoring methods.

Test Material Selection Process

Time was available to do only cursory study of material in sorting it into training and test bins. A vague, intuitive sense of "plain vanilla" vs. "weird" was used. Several sessions were ruled out as test material because they were unusual: one had an extremely low frequency of "Class A" queries; in another, almost all the queries were just NP's, without verbs. In the sessions accepted as test material, all "Class A" queries were used.

It was strongly suggested that we partition the results into "new word" and "old word" sets, the "new word" set being those queries containing words not in the training material. This motivated us to think some about the "new word" problem. Probably the principle being implicitly addressed here is that test queries are unfair if they are not answerable by the logical generalization of the

conjunction of training material and the initial state of the language model. (In a spelling bee, it is probably unfair to expect a contestant to come up with the "k" in "knight" if that word -- or a related word -- had never been seen in training.) This is the other side of the usual constraint between testing and training: that they be statistically independent, for a valid test.

Violation of the "fair generalization" constraint between training set and test set does not make a test "invalid", or necessarily biased, but only inefficient and "unfair" if only the bottom line is paid attention to.

Since some words are understandable even though one has not previously heard them, and polysemous words are not necessarily understandable after limited exposure, "words" are not the right unit to look at in deciding if a test case is implied by the training set. The real constraint is that all sound-to-meaning mappings that are required in order to answer the test question be learnable from the training material (assuming an initial "tabula rasa" language model). This points to idioms: sound-meaning pairs that are not predictable by general rule from a knowledge of the sounds and meanings of their constituent parts. Knowing the meaning of "time" and "table" does not make "time table" understandable. And morphemes (roughly non-complex words) qualify as idioms, since they have no sound-meaning constituent parts. New syntactic constructions mediating between sound and meaning would also make the sound-to-meaning mapping of a query unhandleable.

Here are the qualitatively new elements that we found in the test material (including non-Class A) utterances :

1. Morphemes: EARLY, EQUAL, ITS, LOCKHEED, NIGHT, [STAYING] OVER, and PART.

2. Words: [U] A'S, LEAVES, MEANINGS, MORNINGS, NINETEENTH, PRICES, SEATINGS, SERVICING, [TO] SERVICE, SPECIALS, STAYING [OVER], and THREE'S.

3. Multi-word Idioms: TIME TABLE (only in Class X)

With one exception, none of the five Class A queries with new morphemes were answered correctly. Query bp00k1s, with "NIGHT" in it, was successfully answered by only the MIT system. Perhaps because "NIGHT" is in the knowledge database, it should have been counted as an "old" morpheme.

Several of the ten queries with new complex words in the test set were answered correctly; primarily ones with new words that are regular morphological variants of other words that are in the training set (or the assumed pre-existing language model), e.g. "meanings", "times", or "nineteenth".

A table showing the "new phenomena" subset of results is provided at the end of this paper (Table 2).

It seems to us that a promising research topic would be further study of such training-test "fair generalization" or "learnability" constraints, with an eye to automating detection of their violation in the design of better tests.

Preliminary Results

Results were reported to NIST for a total of seven systems by June 19th: two systems from BBN, two from CMU and one each from MIT, SRI and Unisys. The system designated as cmu-spi ("spi" => speech input) was the only one for which the input consisted of the speech waveform. For the other systems the input consisted of the SNOR transcriptions. Subsequently, reformatted results for three systems were accepted: "cmu-r", "cmu-spir", and "mit-r".

The CAS-format input provided for an answer of the form NO_ANSWER to indicate that the system failed to provide an answer for any of several reasons (e.g., failure to recognize the words, failure to parse, failure to produce a valid database query, etc). Some sites made considerable use of this option, others (e.g., MIT) initially did not, partially due to miscommunication about the validity of this option.

Some trivial fixes in the format used for submissions of results from some of the sites were made. One site initially omitted an answer for one of the queries, throwing subsequent REF-HYP alignment off; we inserted for them a NO_ANSWER response. Since there was miscommunication about the use of the "NO_ANSWER" response, we also changed one system's stock response meaning "system can't handle it" to "NO_ANSWER" for them, and allowed another site to submit revised results with "NO_ANSWER" in place of some of their responses. In the table of results, the revised systems are "cmu-r", "cmu-spir", and "mit-r".

Responding to several complaints from sites about specific items in the test reference material, we corrected one reference answer (bd0071s) and changed the classification of three queries (bm0011s, bp0081s, and bw00s1s) from Class A to Class X (in effect deleting these from the test set, reducing the test set size to 90 valid Class A queries). The classification disputes all centered on ambiguity, one of the hardest calls to make. If similar limitations on what is evaluable are made for the next round, we would like to have both an explicit principle for deciding when ambiguity is present and a procedure for adjudicating disputes agreed on early. The detailed results are given in Table 1a for Class A queries with only lexical items that appear at least once in the training data, and in Table 2a for Class A queries with "new" morphemes, words, or idioms. Table 3 presents a complete summary of the results for the entire 90 sentence-utterance test set.

Since the Class A test queries are not context-dependent, the ordering of these queries is not significant. As an aid in analysis, for the results presented in Tables 1a and 2a, queries have been (roughly) rank ordered in order of increasing apparent difficulty. Note that queries toward the top of both parts of the table resulted in more "T" answers than "F" or "NA", while queries toward the bottom of the table resulted in more "F" and "NA" answers. Not surprisingly, there appears to be a general trend toward increasing apparent difficulty with increased length of the utterance (number of words).

Table 3 shows that the number of correct answers from the various systems ranged from 25 to 58. Note also that for the system for which speech waveform data was used as input, (cmu-spir), 35 of the queries were answered correctly. Comparing results from similar systems for the two subsets of the data (Tables 1b and 2b), note that the ratios of the numbers of correctly recognized queries in the two subsets vary from 1.9 to 4.6, with better performance on the subset for which all lexical items occurred at least once in the training data, of course.

Comparisons such as these are complicated, however, by the fact that different systems returned NO_ANSWER for from 0 to 60 of the queries. Perhaps a more appropriate denominator to be used in computing the percentage of correct responses would have been the number of responses for which an answer was provided.

Summary

This paper has presented results of the first Spoken Language System tests conducted in the DARPA Air Travel Information System domain.

bbn-harc T	bbn-par T	cmu T	cmu-r T	cmu-spi F	cmu-spir F	mit T	mit-r T	sri T	unisys T	8	2	0
b4005lxsx SHOW ME THE DISTANCE FROM THE DENVER AIRPORT TO DOWNTOWN	bbn-par T	cmu T	cmu-r T	cmu-spi F	cmu-spir F	mit T	mit-r T	sri NA	unisys NA	7	0	3
bbn-harc NA	bbn-par T	cmu T	cmu-r T	cmu-spi T	cmu-spir T	mit T	mit-r T	sri NA	unisys NA	7	1	2
b4004lxsx SHOW ME ALL THE FLIGHTS FROM DENVER TO SAN FRANCISCO BETWEEN TWO P M AND SEVEN P M	bbn-par T	cmu T	cmu-r T	cmu-spi F	cmu-spir NA	mit T	mit-r T	sri NA	unisys T	7	1	2
bbn-harc T	bbn-par T	cmu T	cmu-r T	cmu-spi F	cmu-spir NA	mit T	mit-r T	sri NA	unisys T	7	1	2
b4000lxsx WHAT IS THE NAME OF THE AIRPORT IN ATLANTA	bbn-par T	cmu T	cmu-r T	cmu-spi F	cmu-spir NA	mit T	mit-r T	sri NA	unisys T	7	2	1
bbn-harc T	bbn-par T	cmu T	cmu-r T	cmu-spi F	cmu-spir NA	mit T	mit-r T	sri NA	unisys T	7	2	1
b4000lxsx WHAT DOES E Q P STAND FOR	bbn-par T	cmu T	cmu-r T	cmu-spi F	cmu-spir NA	mit T	mit-r T	sri NA	unisys T	7	2	1
bbn-harc T	bbn-par T	cmu T	cmu-r T	cmu-spi F	cmu-spir NA	mit T	mit-r T	sri NA	unisys T	7	2	1
b4000lxsx WHAT IS THE NAME OF THE AIRPORT IN SAN FRANCISCO	bbn-par T	cmu T	cmu-r T	cmu-spi F	cmu-spir NA	mit T	mit-r T	sri NA	unisys T	7	2	1
bbn-harc T	bbn-par T	cmu T	cmu-r T	cmu-spi F	cmu-spir NA	mit T	mit-r T	sri NA	unisys T	7	2	1
b4000lxsx SHOW ME ALL FLIGHT INFORMATION FROM ATLANTA TO SAN FRANCISCO	bbn-par T	cmu T	cmu-r T	cmu-spi F	cmu-spir F	mit T	mit-r T	sri F	unisys T	7	3	0
bbn-harc T	bbn-par T	cmu T	cmu-r T	cmu-spi F	cmu-spir F	mit T	mit-r T	sri F	unisys T	7	3	0
b4000lxsx SHOW ME THE TRANSPORTATION FROM S F O TO DOWNTOWN SAN FRANCISCO	bbn-par T	cmu T	cmu-r T	cmu-spi T	cmu-spir T	mit T	mit-r T	sri NA	unisys NA	6	0	4
bbn-harc NA	bbn-par NA	cmu T	cmu-r T	cmu-spi T	cmu-spir T	mit T	mit-r T	sri NA	unisys NA	6	0	4
b4001lxsx WHAT TYPE OF GROUND TRANSPORTATION IS AVAILABLE FROM D F W AIRPORT TO DOWNTOWN DALLAS	bbn-par T	cmu T	cmu-r T	cmu-spi F	cmu-spir F	mit T	mit-r T	sri NA	unisys F	6	3	1
bbn-harc T	bbn-par T	cmu T	cmu-r T	cmu-spi F	cmu-spir F	mit T	mit-r T	sri NA	unisys F	6	3	1
b4000lxsx SHOW ME ALL THE FLIGHTS FROM BOSTON TO S F O ON JUNE SECOND	bbn-par T	cmu T	cmu-r F	cmu-spi F	cmu-spir F	mit T	mit-r T	sri T	unisys T	6	4	0
bbn-harc T	bbn-par T	cmu T	cmu-r F	cmu-spi F	cmu-spir F	mit T	mit-r T	sri T	unisys T	6	4	0
b4001lxsx SHOW ME ALL THE FLIGHTS FROM S F O TO D F W ON JUNE NINE	bbn-par T	cmu T	cmu-r F	cmu-spi F	cmu-spir F	mit T	mit-r T	sri T	unisys T	6	4	0
bbn-harc T	bbn-par T	cmu T	cmu-r F	cmu-spi F	cmu-spir F	mit T	mit-r T	sri T	unisys T	6	4	0
b4000lxsx SHOW ME GROUND TRANSPORTATION IN SAN FRANCISCO	bbn-par T	cmu T	cmu-r T	cmu-spi F	cmu-spir F	mit T	mit-r T	sri T	unisys T	6	4	0
bbn-harc T	bbn-par T	cmu T	cmu-r T	cmu-spi F	cmu-spir F	mit T	mit-r T	sri T	unisys T	6	4	0
b4000lxsx SHOW ME ALL THE NONSTOP FLIGHTS FROM DALLAS TO SAN FRANCISCO	bbn-par T	cmu T	cmu-r F	cmu-spi F	cmu-spir F	mit T	mit-r T	sri T	unisys T	6	4	0
bbn-harc T	bbn-par T	cmu T	cmu-r F	cmu-spi F	cmu-spir F	mit T	mit-r T	sri T	unisys T	6	4	0
b4000lxsx SHOW ME ALL THE FLIGHTS FROM DALLAS TO SAN FRANCISCO	bbn-par T	cmu T	cmu-r F	cmu-spi F	cmu-spir F	mit T	mit-r T	sri T	unisys T	6	4	0
bbn-harc T	bbn-par T	cmu T	cmu-r F	cmu-spi F	cmu-spir F	mit T	mit-r T	sri T	unisys T	6	4	0
b4000lxsx LIST ALL INFORMATION ABOUT FLIGHTS FROM DALLAS FORT WORTH TO ATLANTA	bbn-par T	cmu T	cmu-r T	cmu-spi T	cmu-spir T	mit T	mit-r T	sri NA	unisys T	5	2	3
bbn-harc NA	bbn-par NA	cmu T	cmu-r T	cmu-spi T	cmu-spir T	mit T	mit-r T	sri NA	unisys T	5	2	3
b4000lxsx WHAT IS FLIGHT CODE ONE ZERO TWO FOUR SEVEN	bbn-par T	cmu T	cmu-r T	cmu-spi F	cmu-spir F	mit T	mit-r T	sri NA	unisys T	5	3	2
bbn-harc T	bbn-par T	cmu T	cmu-r F	cmu-spi F	cmu-spir NA	mit T	mit-r T	sri NA	unisys T	5	3	2
b4000lxsx WHAT GROUND TRANSPORTATION IS AVAILABLE FROM THE AIRPORT IN BOSTON TO DOWNTOWN BOSTON	bbn-par T	cmu T	cmu-r T	cmu-spi F	cmu-spir F	mit T	mit-r T	sri NA	unisys T	5	3	2
bbn-harc T	bbn-par T	cmu T	cmu-r T	cmu-spi F	cmu-spir F	mit T	mit-r T	sri NA	unisys T	5	3	2
b4000lxsx SHOW ME ALL THE FLIGHTS FROM PHILADELPHIA TO BOSTON ON MAY TWENTY SIXTH	bbn-par T	cmu T	cmu-r T	cmu-spi F	cmu-spir F	mit T	mit-r T	sri NA	unisys T	5	4	1
bbn-harc T	bbn-par T	cmu T	cmu-r F	cmu-spi F	cmu-spir F	mit T	mit-r T	sri NA	unisys T	5	4	1
b4000lxsx WHAT IS TRANSPORT L	bbn-par T	cmu T	cmu-r F	cmu-spi F	cmu-spir F	mit T	mit-r T	sri NA	unisys T	5	4	1
bbn-harc T	bbn-par T	cmu T	cmu-r F	cmu-spi F	cmu-spir F	mit T	mit-r T	sri NA	unisys T	5	4	1
b4000lxsx WHAT IS TRANSPORT R	bbn-par T	cmu T	cmu-r F	cmu-spi F	cmu-spir F	mit T	mit-r T	sri NA	unisys T	5	4	1
bbn-harc T	bbn-par T	cmu T	cmu-r F	cmu-spi F	cmu-spir F	mit T	mit-r T	sri NA	unisys T	5	4	1
b4000lxsx WHAT IS TRANSPORT R	bbn-par T	cmu T	cmu-r F	cmu-spi F	cmu-spir F	mit T	mit-r T	sri NA	unisys T	5	4	1
bbn-harc T	bbn-par T	cmu T	cmu-r F	cmu-spi F	cmu-spir F	mit T	mit-r T	sri NA	unisys T	5	4	1
b4000lxsx WHAT IS TRANSPORT L	bbn-par T	cmu T	cmu-r F	cmu-spi F	cmu-spir F	mit T	mit-r T	sri NA	unisys T	5	4	1
bbn-harc T	bbn-par T	cmu T	cmu-r F	cmu-spi F	cmu-spir F	mit T	mit-r T	sri NA	unisys T	5	4	1
b40007lxsx SHOW ME THE NONSTOP FLIGHTS FROM SAN FRANCISCO TO D F W BEFORE NOON	bbn-par T	cmu T	cmu-r F	cmu-spi F	cmu-spir F	mit T	mit-r T	sri T	unisys F	5	5	0
bbn-harc T	bbn-par T	cmu T	cmu-r F	cmu-spi F	cmu-spir F	mit T	mit-r T	sri T	unisys F	5	5	0
b40006lxsx WHAT DOES D L MEAN	bbn-par T	cmu T	cmu-r F	cmu-spi F	cmu-spir F	mit T	mit-r T	sri F	unisys T	5	5	0
bbn-harc T	bbn-par T	cmu T	cmu-r F	cmu-spi F	cmu-spir F	mit T	mit-r T	sri F	unisys T	5	5	0
b4000d1xsx ARE THERE ANY EXCURSION FARES FOR ROUND-TRIP TICKETS FROM DALLAS TO BOSTON	bbn-par NA	cmu T	cmu-r F	cmu-spi T	cmu-spir T	mit T	mit-r T	sri NA	unisys NA	4	2	4
bbn-harc NA	bbn-par NA	cmu T	cmu-r F	cmu-spi T	cmu-spir T	mit T	mit-r T	sri NA	unisys NA	4	2	4
b40006lxsx SHOW ME THE DISTANCE FROM SAN FRANCISCO AIRPORT TO DOWNTOWN	bbn-par F	cmu T	cmu-r T	cmu-spi T	cmu-spir T	mit T	mit-r T	sri NA	unisys NA	4	3	3
bbn-harc NA	bbn-par F	cmu T	cmu-r T	cmu-spi T	cmu-spir T	mit T	mit-r T	sri NA	unisys NA	4	3	3
b40001lxsx WHAT AIRLINE IS M L	bbn-par T	cmu T	cmu-r NA	cmu-spi F	cmu-spir NA	mit T	mit-r T	sri NA	unisys F	4	3	3
bbn-harc T	bbn-par T	cmu T	cmu-r NA	cmu-spi F	cmu-spir NA	mit T	mit-r T	sri NA	unisys F	4	3	3
b40002lxsx WHAT AIRLINE IS C O	bbn-par T	cmu T	cmu-r F	cmu-spi F	cmu-spir F	mit T	mit-r T	sri NA	unisys F	4	5	1
bbn-harc T	bbn-par T	cmu T	cmu-r F	cmu-spi F	cmu-spir F	mit T	mit-r T	sri NA	unisys F	4	5	1
b40007lxsx WHAT DOES L H MEAN	bbn-par T	cmu T	cmu-r NA	cmu-spi F	cmu-spir F	mit T	mit-r T	sri F	unisys F	4	5	1
bbn-harc T	bbn-par T	cmu T	cmu-r NA	cmu-spi F	cmu-spir F	mit T	mit-r T	sri F	unisys F	4	5	1
b40000lxsx WHAT DO THE TRANSPORT ABBREVIATIONS MEAN	bbn-par F	cmu T	cmu-r F	cmu-spi F	cmu-spir NA	mit T	mit-r T	sri NA	unisys NA	3	4	3
bbn-harc T	bbn-par F	cmu T	cmu-r F	cmu-spi F	cmu-spir NA	mit T	mit-r T	sri NA	unisys NA	3	4	3
b4000c1xsx WHAT DOES F A MEAN	bbn-par T	cmu T	cmu-r F	cmu-spi F	cmu-spir NA	mit T	mit-r T	sri NA	unisys T	3	5	2
bbn-harc T	bbn-par T	cmu T	cmu-r F	cmu-spi F	cmu-spir NA	mit T	mit-r T	sri NA	unisys T	3	5	2
b40000lxsx SHOW ME GROUND TRANSPORTATION TYPES IN ATLANTA	bbn-par T	cmu T	cmu-r T	cmu-spi F	cmu-spir NA	mit T	mit-r T	sri NA	unisys F	2	3	5
bbn-harc NA	bbn-par NA	cmu T	cmu-r T	cmu-spi F	cmu-spir NA	mit T	mit-r T	sri NA	unisys F	2	3	5
b4000f1xsx WHAT TYPE OF AIRCRAFT IS FLYING UNITED AIRLINES FLIGHT NINE FIFTY THREE												

Table 1a - contd.

bbn-harc NA bbn-par NA cmu T cmu-spi F cmu-spir F mit F mit-r F sri NA unisys NA 2 4 4
 bm005lhx WHAT IS THE AIRFARE FOR A FIRST CLASS ROUND-TRIP TICKET FROM DALLAS TO BOSTON LEAVING ON DELTA FLIGHT SIX NINETY FOUR RETU
 RNING ON DELTA FLIGHT FIVE EIGHTY NINE cmu T cmu-spi F cmu-spir F mit F mit-r F sri NA unisys NA 2 4 4
 bbn-harc NA bbn-par NA cmu T cmu-spi F cmu-spir F mit F mit-r F sri NA unisys NA 2 4 4
 bm00blsx GIVE ME A LIST OF ADVANCE PURCHASE AIRFARES FOR ROUND-TRIP TICKETS FROM DALLAS TO BOSTON ON DELTA FLIGHT SIX NINETY FOUR R
 ETURNING ON DELTA FLIGHT FIVE EIGHTY NINE cmu F cmu-spi T cmu-spir T mit F mit-r F sri NA unisys NA 2 4 4
 bp00clsx SHOW ME A LIST OF CODES FOR THE MEALS cmu F cmu-spi T cmu-spir T mit F mit-r F sri NA unisys NA 2 4 4
 bbn-harc T bbn-par T cmu F cmu-spi F cmu-spir NA mit F mit-r F sri NA unisys NA 2 4 4
 bp00zlsx IS FIVE HUNDRED AND FIFTY TWO DOLLARS THE CHEAPEST FARE FROM SAN FRANCISCO TO DALLAS FORT WORTH cmu F cmu-spi F cmu-spir T mit F mit-r F sri NA unisys NA 2 4 4
 bbn-harc NA bbn-par NA cmu F cmu-spi T cmu-spir T mit F mit-r F sri NA unisys NA 2 4 4
 bw009lhx WHAT TYPES OF GROUND TRANSPORTATION SERVICES ARE AVAILABLE FROM THE AIRPORT IN ATLANTA TO DOWNTOWN ATLANTA cmu T cmu-spi F cmu-spir F mit F mit-r F sri NA unisys F 2 4 4
 bbn-harc NA bbn-par NA cmu T cmu-spi F cmu-spir F mit F mit-r F sri NA unisys F 2 4 4
 bm00elx GIVE ME A LIST OF FLIGHTS FROM DALLAS TO BOSTON THAT ONLY HAVE ONE STOP BETWEEN DALLAS AND BOSTON cmu F cmu-spi T cmu-spir T mit F mit-r F sri NA unisys F 2 5 3
 bbn-harc NA bbn-par NA cmu F cmu-spi T cmu-spir T mit F mit-r F sri NA unisys NA 2 5 3
 bp00rlsx SHOW ME THE FARES FOR EACH TYPE OF GROUND TRANSPORTATION IN ATLANTA cmu F cmu-spi F cmu-spir NA mit F mit-r F sri NA unisys T 2 6 2
 bbn-harc T bbn-par T cmu F cmu-spi F cmu-spir A mit F mit-r F sri NA unisys T 1 6 3
 bw00alx UNDER THE CATEGORY GROUND TRANSPORTATION WHAT IS TRANSPORT A cmu F cmu-spi F cmu-spir F mit F mit-r F sri NA unisys T 1 6 3
 bbn-harc NA bbn-par T cmu F cmu-spi F cmu-spir F mit F mit-r F sri NA unisys T 1 6 3
 bm00alx WHAT DOES THE D L UNDER THE COLUMN F A MEAN cmu F cmu-spi F cmu-spir F mit F mit-r F sri NA unisys NA 0 4 6
 bp00mlsx SHOW ME FLIGHT NUMBERS OF AMERICAN FROM DALLAS FORT WORTH TO ATLANTA cmu F cmu-spi F cmu-spir NA mit F mit-r F sri NA unisys NA 0 5 5
 bbn-harc NA bbn-par NA cmu F cmu-spi F cmu-spir NA mit F mit-r F sri NA unisys NA 0 5 5
 bm004lhx I NEED FLIGHT TIMES FROM BOSTON TO DALLAS LEAVING ON SUNDAY AFTERNOON AFTER THREE O'CLOCK cmu F cmu-spi F cmu-spir NA mit F mit-r F sri NA unisys NA 0 5 5
 bbn-harc NA bbn-par NA cmu F cmu-spi F cmu-spir NA mit F mit-r F sri NA unisys NA 0 5 5
 bm007lhx GIVE ME A LIST OF ALL AIRFARES FOR A ROUND-TRIP TICKETS FROM DALLAS TO BOSTON ON DELTA FLIGHT SIX NINETY FOUR RETURNING ON
 DELTA FLIGHT FIVE EIGHTY NINE cmu F cmu-spi F cmu-spir NA mit F mit-r F sri NA unisys NA 0 5 5
 bbn-harc NA bbn-par NA cmu F cmu-spi F cmu-spir NA mit F mit-r F sri NA unisys NA 0 5 5
 bm00hlx GIVE ME A LIST OF ALL AIRFARES FOR ROUND-TRIP TICKETS FROM DALLAS TO BOSTON FLYING ON AMERICAN AIRLINES cmu F cmu-spi F cmu-spir F mit F mit-r F sri NA unisys NA 0 5 5
 bbn-harc NA bbn-par NA cmu F cmu-spi F cmu-spir F mit F mit-r F sri NA unisys NA 0 5 5
 bm00jlsx GIVE ME A LIST OF ALL AIRFARES FOR ROUND-TRIP TICKETS FLYING FROM DALLAS TO PHILADELPHIA ON AMERICAN AIRLINES cmu F cmu-spi F cmu-spir F mit F mit-r F sri NA unisys NA 0 5 5
 bbn-harc NA bbn-par NA cmu F cmu-spi F cmu-spir F mit F mit-r F sri NA unisys NA 0 5 5
 bm00klsx GIVE ME A LIST OF ALL AIRFARES FOR ROUND-TRIP TICKETS FLYING FROM DALLAS TO SAN FRANCISCO ON AMERICAN AIRLINES cmu F cmu-spi F cmu-spir F mit F mit-r F sri NA unisys NA 0 5 5
 bbn-harc NA bbn-par NA cmu F cmu-spi F cmu-spir F mit F mit-r F sri NA unisys NA 0 5 5
 bp012lsx SHOW ME ALL INFORMATION ABOUT THE AIRCRAFT TYPE BOEING SEVEN SIX SEVEN cmu NA cmu-spi F cmu-spir NA mit F mit-r F sri NA unisys F 0 5 5
 bbn-harc NA bbn-par F cmu NA cmu-spi F cmu-spir NA mit F mit-r F sri NA unisys NA 0 6 4
 bm00clsx ARE THERE ANY ADVANCE PURCHASE FARES FROM DALLAS TO BOSTON FOR ROUND-TRIP TICKETS cmu F cmu-spi F cmu-spir F mit F mit-r F sri NA unisys NA 0 7 3
 bbn-harc NA bbn-par NA cmu F cmu-spi F cmu-spir F mit F mit-r F sri NA unisys NA 0 7 3
 bf00dlx SHOW ME ALL THE FLIGHTS AND THEIR FARES FROM SAN FRANCISCO TO BOSTON ON JUNE SECOND cmu F cmu-spi F cmu-spir F mit F mit-r F sri NA unisys NA 0 7 3
 bbn-harc NA bbn-par F cmu F cmu-spi F cmu-spir F mit F mit-r F sri NA unisys NA 0 7 3
 bp00hlx HOW MUCH DOES THE FLIGHT FROM DALLAS FORT WORTH TO ATLANTA ROUND-TRIP COST cmu F cmu-spi F cmu-spir F mit F mit-r F sri NA unisys NA 0 7 3
 bbn-harc NA bbn-par F cmu F cmu-spi F cmu-spir F mit F mit-r F sri NA unisys NA 0 7 3

Table 1a - contd.

Site counts	T	F	NA
bbn-harc	50	0	26
bbn-par	53	5	18
cmu	41	34	1
cmu-r	41	31	4
cmu-spi	33	43	0
cmu-spir	33	29	14
mit	51	25	0
mit-r	51	18	7
sri	23	4	49
unisys	45	10	21

Summarized ATIS Results for Utterances with Lexical Items Contained in Training Lexicon

where T is for a correct answer agreement of REF and HYP
 F is for an incorrect answer
 NA is for "NO_ANSWER"

Table 1b.

Detailed ATIS Results for Utterances with Lexical Items Not Contained in Training Lexicon

	T	F	NA
bf002lxs SHOW ME THE PRICES FOR FLIGHTS FROM DALLAS TO BALTIMORE ON MAY TWELFTH	mit	T	
bbn-harc NA bbn-par T cmu T cmu-i T cmu-spi T cmu-spir T	mit	T	
bd001lxs SHOW ME THE FLIGHT THAT LEAVES SAN FRANCISCO FOR D F W AT NINE A M	mit	T	
bbn-harc T bbn-par T cmu T cmu-i T cmu-spi T cmu-spir T	mit	T	
bf008lxs SHOW ME ALL THE FLIGHTS FROM BALTIMORE TO PHILADELPHIA ON MAY NINETEENTH	mit	T	
bbn-harc T bbn-par T cmu F cmu-i F cmu-spi F cmu-spir F	mit	T	
bd00hlxs SHOW ME WHICH AIRLINE FLIGHT LEAVES FROM DENVER TO SAN FRANCISCO AT EIGHTEEN TEN	mit	T	
bbn-harc NA bbn-par T cmu T cmu-i T cmu-spi F cmu-spir NA	mit	F	
bp00klxs DO YOU HAVE TO TAKE A Y N FLIGHT ONLY AT NIGHT	mit	F	
bbn-harc NA bbn-par NA cmu F cmu-i NA cmu-spi F cmu-spir NA	mit	T	
bm002lxs IS THERE MORE THAN ONE AIRPORT IN THE BOSTON AREA THAT AMERICAN AND DELTA SERVICE	mit	T	
bbn-harc NA bbn-par NA cmu F cmu-i NA cmu-spi F cmu-spir NA	mit	T	
bd00klxs SHOW ME THE AIRFARES ON U A'S FLIGHT FIVE ONE ONE	mit	T	
bbn-harc NA bbn-par T cmu F cmu-i F cmu-spi F cmu-spir NA	mit	F	
bm003lxs PLEASE GIVE ME A LIST OF FLIGHT TIMES FROM DALLAS TO BOSTON LEAVING ON SATURDAY MORNINGS BEFORE NOON	mit	F	
bbn-harc NA bbn-par NA cmu F cmu-i NA cmu-spi F cmu-spir NA	mit	F	
bp00alsx LIST THE DAYS AND ITS MEANINGS	mit	F	
bbn-harc NA bbn-par NA cmu F cmu-i NA cmu-spi F cmu-spir NA	mit	F	
bp01llxs SHOW ME ALL INFORMATION ABOUT AIRCRAFT TYPE LOCKHEED L ONE ZERO ONE ONE	mit	F	
bbn-harc NA bbn-par NA cmu F cmu-i NA cmu-spi F cmu-spir NA	mit	F	
bd001lxs SHOW ME ALL THE NONSTOP FLIGHTS FROM DALLAS TO DENVER EARLY IN THE MORNING	mit	F	
bbn-harc NA bbn-par NA cmu F cmu-i F cmu-spi F cmu-spir NA	mit	F	
bm001lxs I NEED INFORMATION ON AIRLINES SERVICING BOSTON FLYING FROM DALLAS	mit	F	
bbn-harc NA bbn-par NA cmu F cmu-i F cmu-spi F cmu-spir F	mit	F	
bp009lxs ON THE DAYS IS A ONE EQUAL TO SUNDAY	mit	F	
bbn-harc NA bbn-par NA cmu F cmu-i NA cmu-spi F cmu-spir F	mit	F	
bp013lxs SHOW ME CAPACITY SEATINGS FOR THE BOEING SEVEN SIXTY SEVEN	mit	F	
bbn-harc NA bbn-par F cmu F cmu-i F cmu-spi F cmu-spir F	mit	F	

Table 2a.

Summarized ATIS Results for Utterances with Lexical Items Not Contained in Training Lexicon

Site counts	T	F	NA
bbn-harc	2	0	12
bbn-par	5	1	8
cmu	3	11	0
cmu-i	3	5	6
cmu-spi	2	12	0
cmu-spir	2	4	8
mit	5	9	0
mit-r	4	7	3
sri	2	1	11
unisys	2	1	11

Where T is for a correct answer agreement of REF and HYP
 F is for an incorrect answer
 NA is for "NO_ANSWER"

Table 2b.

Summarized ATIS Results for All Utterances

Site counts	T	F	NA
bbn-harc	52	0	38
bbn-par	58	6	26
cmu	44	45	1
cmu-i	44	36	10
cmu-spi	35	55	0
cmu-spir	35	33	22
mit	56	34	0
mit-r	55	25	10
sri	25	5	60
unisys	47	11	32

Where T is for a correct answer agreement of REF and HYP
 F is for an incorrect answer
 NA is for "NO_ANSWER"

Table 3.