APPENDIX B:

TEST PROCEDURES

1. GENERAL INSTRUCTIONS

Testing may be done any time during the week of 6-12 May. The only requirement is that all reports (see section 4, below) be received by NOSC by first thing Monday morning, 13 May. Permission to attend MUC-3 at NOSC on 21-23 May may be revoked if you do not meet this deadline!

To complete the required testing, you will need approximately the same amount of time as it would normally take you to run 100 texts in DEV and interactively score them, plus some time to permit you to be extra careful doing the interactive scoring (since the resulting history file is to be used for all passes through the scoring program) and some time for the initializations of the scoring program with the different configuration files required for the various linguistic phenomena tests. If you carry out the optional testing, you will need to allow time to generate at least a couple new sets of response templates. In that case, you will also need time to add to the history file as needed during the additional scoring runs.

IF YOU INTEND TO CARRY OUT ANY OF THE OPTIONAL TESTING, YOU MUST REPORT THE PLANNED "PARAMETER SETTINGS" TO NOSC FOR BOTH THE REQUIRED TEST AND THE OPTIONAL TESTING BEFORE STARTING THE TEST PROCEDURE. This means that you should describe, in some meaningful terms, SPECIFICALLY how you will alter the behavior of the system so that it will produce each of the different tradeoffs in metrics described in the sections below.

1.1 REQUIRED TESTING: MAXIMIZED RECALL/PRECISION TRADEOFF

To ensure comparability among the test results for all systems, THE REQUIRED TESTING MUST BE CONDUCTED WITH THE SYSTEM SET TO MAXIMIZE THE TRADEOFF BETWEEN RECALL AND PRECISION IN THE MATCHED/MISSING ROW IN THE SCORE SUMMARY REPORT. The maximum of recall and precision does not mean an ADDITIVE maximization, but that the total scores for each of the two metrics should be as close together and as high as possible. For most systems, this is probably the normal way the system operates.

Several passes through the scoring program will be required, one for the official test on generating templates for the whole test set and the others for the experimental tests on generating the specific slots called out by the linguistic phenomena tests. You generate only one set of system responses, and only the first pass through the scoring program will require user interaction. The history file produced during this interaction will be used in the scoring of the linguistic phenomena tests. (It will also serve as the basis for scoring any optional tests that are conducted.)

1.2 OPTIONAL TESTING: OTHER RECALL/PRECISION TRADEOFFS

The objective of the optional testing is to learn more about the tradeoffs that some systems may be designed to make between recall and precision. It is intended to elicit extra data points only on those systems that are currently designed to make some theoretically interesting tradeoffs in some controlled fashion.

Thus, we are interested in having you conduct the optional testing in either of the two following cases, but not otherwise:

1) if the system can control the tradeoff between recall and precision in order to produce a set of data points sufficient to plot the outline of a recall-precision curve;

2) if the system's recall and precision can be consciously manipulated by the loosening or tightening of analysis constraints, etc., in order to produce at least one data point that contrasts in an interesting way with the results produced by the required testing.

To yield these additional data points, you will generate and score new system response templates, using the history file generated during the required testing. NO SYSTEM DEVELOPMENT IS PERMITTED BETWEEN OFFICIAL TESTING AND OPTIONAL TESTING -- ONLY MODIFICATION OF SYSTEM CONTROL PARAMETERS AND/OR REINSERTION OR DELETION OF EXISTING CODE THAT AFFECTS THE SYSTEM'S BEHAVIOR WITH RESPECT TO THE TRADEOFF BETWEEN RECALL AND PRECISION.

If, as a consequence of altering the system's behavior, templates are generated that weren't generated during the required testing or slots are filled differently, you may find it necessary to add to the history file and to change some of the manual template remappings. START THE SCORING OF EACH OPTIONAL TEST WITH THE HISTORY FILE GENERATED DURING THE REQUIRED TESTING, MINUS THE MANUAL TEMPLATE REMAPPINGS; SAVE ANY UPDATED HISTORIES TO NEW FILE NAMES.

In order to obtain these data points, you may wish to conduct a number of tests and throw out all but the best ones. Remember, however, that you are to notify NOSC of ALL the planned parameter settings in advance (see section 1). Thus, it would be wise to experiment on the training data and use the results to know what different runs are worth making during the test. If, among the "throwaways" there are some results that you find significant, you may wish to include them in your site report for the MUC-3 proceedings, but they will not be part of the official record.

You may submit results for the experimental linguistic phenomena tests as part of the optional testing if you wish, but please do so only if you find the differences in scores to be significant.

2. SPECIFIC PROCEDURES FOR THE REQUIRED TESTING

2.1 FREEZING THE SYSTEM AND FTP'ING THE TEST PACKAGE

When you are ready to run the test, ftp the files in the test package from /pub/tst2. You are on your honor not to do this until you have completely frozen your system and are ready to conduct the test. You must stop all system development once you have ftp'ed the test package.

Note: If you expect to be running the test over the weekend and are concerned that a host or network problem might interfere with your ability to ftp, you may ftp the files on Friday. However, for your own sake, minimize the accessibility of those files, e.g., put them in a protected directory of someone who is not directly involved in system development.

2.2 GENERATING THE SYSTEM RESPONSE TEMPLATES

There are 100 texts in tst2-muc3, and the message IDs have the following format: TST2-MUC3-nnnn. Without looking at the texts, run your system against the file and name the output file response-max-tradeoff.tst2.

You are to run the required test only once -- you are not permitted to make any changes to your system until the test is completed. If you get part way through the test and get an error that requires user intervention, you may intervene only to the extent that you are able to continue processing with the NEXT message. You are not allowed to back up!

Notes:

1) If you run short on time and wish to break up tst2-muc3 and run portions of it in parallel, that's fine as long as you are truly running in parallel with a single system or can completely simulate a parallel environment, i.e., the systems are identically configured. You must also be sure to concatenate the outputs before submitting them to the scoring program.

2) No debugging of linguistic capability can be done when the system breaks. For example, if your system breaks when it encounters an unknown word and your only option for a graceful recovery is to define the word, then abort processing and start it up again on the next test message.

3) If you get an error that requires that you reboot the system, you may do so, but you must pick up processing with the message FOLLOWING the one that was being processed when the error occurred. If, in order to pick up processing at that point, you need to create a new version of tst2-muc3 that excludes the messages already processed or you need to start a new output file, that's ok. Be sure to concatenate the output files before submitting them to the scoring program.

2.3 SCORING THE SYSTEM RESPONSE TEMPLATES

2.3.1 SCORING ALL SYSTEM RESPONSES FOR OFFICIAL, REQUIRED TEST

Run the scoring program on the system response templates, using key-tst2 as the answer key and entering config.el as the argument to initialize-muc-scorer. (The config file contains arguments to the define-muc-configuration-options function, which you will have to edit to supply the proper pathnames). When you enter the scoring program, type "1s" so that the score buffer will contain detail tables (template by template) as well as the final summary table. Save the score buffer (*MUC Score Display*) to a file called scores-max-tradeoff.tst2.

Note: During the interactive scoring, make use of the guidelines (supplied separately) for interactively assigning full and partial credit. Also refer to key-tst2-notes (in the ftp directory) for NOSC's comments on how the answer key was generated. See section 5, below, for information on the plans for handling the rescoring of results.

Following the instructions in the user manual for the scoring program, save the history to a file called history-max-tradeoff.tst2.

2.3.2 SCORING SPECIFIC SETS OF SLOTS FOR THE EXPERIMENTAL, REQUIRED LINGUISTIC PHENOMENA TESTS

Read the file readme.phentest. Run the scoring program again for each of the linguistic phenomena tests, i.e., type the configuration file names that appear in the test package in sequence as the argument to the function initialize-muc-scorer. (These files must be edited to provide the proper pathnames for your environment.)

Scoring for the phenomena testing should be done using the history file created when all templates were scored. No updates to the history file should be made during these runs. Save each score buffer (*MUC Score Display*) to the file name scores-<phenomenon test name>-max-tradeoff.tst2, where <phenomenon test name> matches the names in the config files.

3. SPECIFIC PROCEDURES FOR OPTIONAL TESTING

3.1 WITH MODIFIED SYSTEM CONTROL PARAMETERS FOR ALL TEMPLATES

For each optional run, modify the system as specified IN ADVANCE to NOSC. Then follow the procedures described in section 1.2 and section 2. Save the system response templates to files with unique, meaningful names. When you do the scoring, start the scoring program each time with the history file generated during the required testing (minus the manually remapped templates, since you may wish to change them), and save the history when you have finished scoring (whether it was updated or not) and the scores to files with names that permit them to be matched up with the corresponding system response template file.

Once you have determined which of the optional runs to submit to NOSC for the official record, name the files for those runs in some meaningful, easily-understood fashion (fitting these patterns: response-<meaningful name here>.tst2, scores-<meaningful name here>.tst2, and history-<meaningful name here>.tst2) and provide them along with a readme file that explains the significance of the files and identifies their corresponding parameter setting.

3.2 FOR LINGUISTIC PHENOMENA TESTS, USING MODIFIED SYSTEM CONTROL PARAMETERS

After you have produced the files listed at the end of section 3.1, above, follow the procedures in section 2.3.2 if you wish to produce separate linguistic phenomena test results for any/all of them. Use the history file corresponding to each of those response files.

Please submit these linguistic phenomena test scores to NOSC only if they are significantly different from those produced for the required testing. If you do submit these scores, name the file for each of the phenomena tests to correspond with the appropriate response file, using the following pattern: scores-<phenomenon test name>-<meaningful name here>.tst2.

4. REPORTS TO BE SUBMITTED TO NOSC BY MONDAY MORNING, MAY 13

All results submitted to NOSC are considered "official," with the exception of the results of the linguistic phenomena testing, which are considered "experimental." All results, whether official or experimental, may be included, in part or in full, in publications resulting from MUC-3. However, only the official results may be used for any comparative ranking or rating of systems. The proper means of using the official results for that purpose will be discussed during the conference at NOSC. The results of the linguistic phenomena testing are to be used only to gain insight into the linguistic performance of individual systems and into the testing methodology.

The files listed below are to be submitted to NOSC by Monday morning, May 13, via email to sundheim@nosc.mil. TO HELP NOSC FILE THE MESSAGES ACCURATELY, PLEASE SUBMIT EACH FILE IN A SEPARATE MESSAGE, AND IDENTIFY YOUR ORGANIZATION AND THE FILE NAME IN THE SUBJECT LINE OF THE MESSAGES.

4.1 REQUIRED TESTING (MAXIMIZED RECALL/PRECISION TRADEOFF)

1. response-max-tradeoff.tst2
2. history-max-tradeoff.tst2
3. scores-max-tradeoff.tst2
4. trace-max-tradeoff.tst2 (system trace for the 100 messages) -- You may submit whatever you think is appropriate, i.e., whatever would serve to help validate the results of testing. If the traces are voluminous and you do not wish to email them, please compress them and ftp them to the /pub directory; send sundheim@nosc.mil an email message to identify the file name.
5. scores-<phenomenon test name>-max-tradeoff.tst2 -- where <phenomenon test name> matches the names in the config files (see readme.phentest)

4.2 OPTIONAL TESTING (OTHER RECALL/PRECISION TRADEOFFS)

Items 1-5, below, are required for EACH optional test run that is reported to NOSC.

1. history-<meaningful name here>.tst2
2. response-<meaningful name here>.tst2
3. scores-<meaningful name here>.tst2
4. readme-optional-testing.tst2 -- See section 3.1, above.
5. trace-<meaningful name here>.tst2 -- See note in section 4.1, above.
6. scores-<phenomenon test name>-<meaningful name here>.tst2 -- where <phenomenon test name> matches the names in the config files (see readme.phentest). Submit these scores only if significantly different from those obtained for the required testing.

5.0 RESCORING OF RESULTS

The interactive scoring that is done during testing should be done in strict conformance to the scoring guidelines. If you perceive errors in the guidelines or in the answer keys as you are doing the scoring, please make note of them and send a summary to NOSC along with the items listed in section 4, above. When all the results are in, NOSC will attempt to merge everyone's history-max-tradeoff.tst2 files and rescore everyone's response-max-tradeoff.tst2 files. Your notes on perceived errors may be useful to NOSC at that time. If the errors are not easy to rectify and if they appear to be serious enough to significantly affect the legitimacy of the scoring, we may have to wait to rectify them after the conference and rescore the response templates at that time. THE RESULTS OF RESCORING BEFORE AND/OR AFTER THE CONFERENCE WILL BECOME THE OFFICIAL RESULTS.