APPENDIX C:

GUIDELINES FOR INTERACTIVE SCORING

1. INTRODUCTION

This document, although fairly extensive, is not intended to give you an exhaustive list of "do's" and "don'ts" about doing the interactive scoring of the templates. Instead, it presents you with guidelines and some examples, in order to imbue you with the spirit of the enterprise. It is up to you to carefully consider your reasons before judging mismatching responses to be "completely" or "partially" correct.

Thus, you should attempt to set aside a substantial amount of time to do the interactive scoring and should plan to do it when you are rested and can be as objective as humanly possible about your system's performance. Please refer to the file key-tst2-notes for examples of decisions NOSC made in preparing the answer key. If you have any doubt whether any given system response deserves to be judged completely/partially correct, count it incorrect.

2. SETTING UP THE SCORING PROGRAM IN INTERACTIVE MODE

You must use the latest official version of the scoring program together with the latest slotconfig.el file. You are not permitted to make any modifications of your own to the scoring software or the files it uses, except to define the pathnames in the config.el file for the files that it reads in.

The configuration (config.el) files supplied with the test package set the :query-verbose option on, which places the scoring program in interactive mode. (See MUC Scoring System User's Manual, section 5.2.) The only feature of the interactive scoring that you are *not* permitted to take advantage of is the option to change a key or response template! This feature is controlled by the :disable-edit option, which is set on in the config.el files supplied in the test package and should not be modified.

Although there may be errors in the key templates, you are not permitted to fix them, as we do not have sufficient time to make the corrections known to all sites. Score your system under the assumption that the answer key is correct, make note of any perceived errors in the key, and email them to NOSC along with your results. If there is sufficient evidence that errors were made that affect the scores obtained, a new key will be prepared after the conference, and sites will be given the opportunity to rescore their system responses. The new scores will replace the old ones as the official results.

Included among your options for interactive scoring is the manual realignment of response templates with key templates (see section 3.2.1 below and section 4.7 of User's Manual). If you are not already comfortable using the interactive scoring features of the scoring program, take some time to practice on some texts in the training set before you attempt to do the scoring for the test set. Also be sure to read the document on test procedures carefully re saving your history buffer to a file for use in other scoring sessions required for completing the test procedure. Reference to key-tst2-notes while you are doing the interactive scoring might help you understand the key better and give you ideas on cases when alternative fillers might be justified.

3. SCORING MISMATCHED SLOT FILLERS

3.1 BY TYPE OF FILL

These subsections deal in turn with string fills, set fills, and other types of fills. Following that is a section concerning cross-reference tags.

3.1.1 STRING FILLS

Slots requiring string fills are slots 5, 6, 8, and 11. In the case of a mismatch on fillers for these slots, the scoring program will permit you to score the response as fully correct, partially correct, or incorrect

3.1.1.1 FULLY CORRECT

NOSC has attempted to provide a choice of good string options for each string slot. If you get a mismatch, before you score a filler fully correct you should consider carefully whether your system's filler is both complete enough and precise enough to show that the system found exactly the right information.

The most likely situation where "fully correct" would be justified is in a case where the system or the key includes "nonessential modifiers" such as articles, quantifiers, and adjectivals for nationalities (e.g., SALVADORAN).

***EXAMPLE** (slot 11):*	*RESPONSE*	*"THE 3 PEASANTS"*
	KEY	*"PEASANTS"*

In filling the key templates, such nonessential modifiers were generally included in slot 5 (since there are no slots specifically for the number and nationality of the perpetrators). They were generally excluded from fillers for the other string slots, unless they seemed to be part of a proper name (e.g. THE EXTRADITABLES).

"Fully correct" is also warranted if the system response contains more modifying words and phrases than the answer key, as long as all the modifiers are modifiers of the noun phrase. However, in most cases the answer key should already contain options such as these.

***EXAMPLE** (slot 11):*	*RESPONSE*	*"OLD PEASANTS WHO WERE WITNESSES"*
	KEY	*"PEASANTS" / "OLD PEASANTS"*

Finally, if your system does not generate an escape (backslash) character in front of the inner double quote marks of a filler that is surrounded by double double quotes, you may score the system response as completely correct if it would otherwise match the key.

EXAMPLE:	*RESPONSE*	*""FOO""*
	KEY	*"\"FOO\"" / "FOO"*

3.1.1.2 PARTIALLY CORRECT

You may score a filler partially correct, but not fully correct, if your system goes overboard and includes adjuncts in the response string that aren't part of the desired noun phrase.

EXAMPLE (slot 11): RESPONSE "THE 3 PEASANTS, WHICH THE GOVERNMENT ADMITTED WAS A MISTAKE"
KEY "PEASANTS"

Scoring a filler partially correct is also appropriate in cases where the key contains a proper name (in the most complete form found in the text) and the response contains only part of the name (i.e., uses an incomplete form found in the text).

EXAMPLE (slot 11): RESPONSE TORRES" ("BODYGUARD")
KEY "ALBERTO ROBERTO TORRES" ("BODYGUARD")

3.1.2 SET FILLS

Slots requiring set fills are slots 3, 4, 7, 10, 13, 14, 15, 17, and 18. (Slot 16, the LOCATION slot, is not treated by the scoring program as having set fills.) In the case of a mismatch on fillers for these slots, the scoring program will not permit you to score them as fully correct. (But see section 3.1.4 below re an exception. Also, see 3.2.7 and 3.2.15 for information concerning automatic assignment of partial credit by the scoring program.)

NOSC has attempted to offer all the possible alternative correct fillers as options in the key; however, scoring a filler partially correct may be justified in certain cases. See the appropriate subsections of section 3.2 below.

3.1.3 OTHER TYPES OF FILLS

Slots requiring other types of fills are slots 1, 2, 9, 12, and 16. In the case of a mismatch on fillers for these slots, the scoring program will permit you to score the fillers as fully correct, partially correct, or incorrect. (But see section 3.1.4 below re an exception. Also, see 3.2.16 for information concerning automatic assignment of partial credit by the scoring program.)

NOSC has attempted to offer all the possible alternative correct fillers as options in the key; however, scoring a filler completely or partially correct may be justified in certain cases. See the appropriate subsections of section 3.2 below.

3.1.4 FILLS THAT INCLUDE CROSS-REFERENCE TAGS

3.1.4.1 FULLY CORRECT

The scoring program permits you to score a slot as fully correct in the case of a mismatch on the slots listed in 3.1.2 and 3.1.3 above where the only mismatch is on a cross-reference tag. In such cases, you may score the entire filler as fully correct only if the filler of the slot indicated by the cross-reference tag was also scored as fully correct.

3.1.4.2 PARTIALLY CORRECT

If the non-tag portion of the filler is not judged completely correct (by the criteria found in other sections of this set of guidelines), the best you can do is to judge the entire filler partially correct. If the non-tag portion is *completely* correct and the tag is either missing or incorrect, it is appropriate to score the entire filler partially correct.

Scoring the entire filler partially correct may also be done if the non-tag portion of the filler is judged *partially* correct and the tag is either missing or incorrect. In this case, however, you must re-read the text and judge the partial correctness of the non-tag portion with respect to the way the text refers to the *KEY'S* tag, not the system response tag. In other words, you must be able to show that the system got the non-tag portion partially correct for the right reason. (Note that this guideline is based on the assumption that some systems might intentionally, not accidentally, generate a correct filler and, for independent reasons, give it an incorrect tag.)

EXAMPLE (slot 7):	*RESPONSE*	*SUSPECTED OR ACCUSED: "RIGHT-WINGERS"*
	KEY	*REPORTED AS FACT: "LEFT-WINGERS"*

*(where SUSPECTED OR ACCUSED has been judged partially correct with respect to its *CORRECT* intended referent, "LEFT-WINGERS", i.e., on the basis of presuming that the whole system response was SUSPECTED OR ACCUSED: "LEFT-WINGERS" rather than SUSPECTED OR ACCUSED: "RIGHT-WINGERS")*

3.1.4.3 INCORRECT

If the non-tag portion of the filler is judged incorrect, then the entire filler must be judged incorrect, even if the tag portion is correct or partially correct.

3.2 BY INDIVIDUAL SLOT

3.2.1 Slot 1 -- TEMPLATE ID

The guidelines here concern the manual realignment of templates in the case where the automatic template mapping facility provided by the scoring program fails to identify the optimal mapping between the set of response templates for a message and the set of key templates for that message. Guidelines are needed because it is possible for the user to elect not to map a response template to any key template at all, i.e., to map a response template to NIL and a key template to NIL rather than mapping the templates to each other. The user may wish to do this in cases where the match between the response and the key is so poor and the number of mismatching fillers so large that the user would rather penalize the system's recall and overgeneration (by mapping to NIL) than penalize the system's precision.

However, to ensure the validity of the performance measures and to ensure comparability among the systems being evaluated, it is important that this option not be overused. The basic rule is that the user must permit a mapping between a response template and a key template if there is a full or partial match on the incident type. (The condition concerning a partial match covers the two basic situations described in section 3.2.3 below.) If there is no match on the incident type, manually mapping to NIL is allowed, at the discretion of the user.

3.2.2 Slot 2 -- DATE OF INCIDENT

FULLY CORRECT OR PARTIALLY CORRECT:

System response is close to the key's date or range of dates (if the date is difficult to calculate). In the example below, the system's response may be judged fully correct, since the system has calculated a more precise date than what was expected by the key.

EXAMPLE:	*TEXT*	*"X OCCURRED ON AUGUST 30, 1989, AND Y OCCURRED A WEEK LATER"*
	RESPONSE	*(for Y) 06 SEP 89*
	KEY	*(for Y) 30 AUG 89 - 15 SEP 89*
	(where the latter date is the date of the article)	

PARTIALLY CORRECT:

1. System response is part of the date contained in the key (either if an incident occurred between two dates or if the filler in the key is a default value, i.e., consists of a range with the date from the message dateline as the upper anchor).

EXAMPLES:	*RESPONSE*	*26 AUG 89*
	KEY	*25 AUG 89 - 26 AUG 89*
	RESPONSE	*26 AUG 89*
	KEY	*26 AUG 89 (default fill)*
	RESPONSE	*25 AUG 89*
	KEY	*-26 AUG 89 (default fill)*

2. System response is a default-looking value (as described above) and the key is a bounded range that has the date of the message dateline as the upper anchor.

EXAMPLE:	*RESPONSE*	*- 26 AUG 89 (default-looking fill)*
	KEY	*25 AUG 89 - 26 AUG 89*

NOTE: The system response should be judged INCORRECT when the response is a default-looking value (as described above) and the key does not have the default anchor date as its value or, in the case of a range, as the upper anchor.

EXAMPLES:	*RESPONSE*	*- 26 AUG 89 (default-looking fill)*
	KEY	*25 AUG 89*
	RESPONSE	*- 26 AUG 89 (default-looking fill)*
	KEY	*24 AUG 89 - 25 AUG 89*

3.2.3 Slot 3 -- TYPE OF INCIDENT

FULLY CORRECT:

Mismatch not allowed to be scored fully correct.

PARTIALLY CORRECT:

The number of cases where it is justifiable to score this slot partially correct should be limited, especially for situations other than the following:

1. System response is the correct incident type, except that ATTEMPTED or THREAT is missing.

2. System response is ATTACK instead of the specific incident type found in the key.

3.2.4 Slot 4 -- CATEGORY OF INCIDENT

FULLY CORRECT:

Mismatch not allowed to be scored fully correct.

PARTIALLY CORRECT:

Since there are only two set fills for this slot, there should be few instances where a mismatch should result in scoring the filler partially correct.

3.2.5 Slot 5 -- PERPETRATOR: ID OF INDIV(S)

FULLY CORRECT:

See section 3.1.1.1.

PARTIALLY CORRECT:

See section 3.1.1.2.

3.2.6 Slot 6 -- PERPETRATOR: ID OF ORG(S)

FULLY CORRECT:

1. In general, the guidelines in section 3.1.1.1 do not apply to this slot, since this slot is intended to be filled only with proper names. However, the term "proper names" is not completely defined, especially with respect to the expected fillers in the case of STATE-SPONSORED TERRORISM. You have more leeway to score fillers as fully correct in such cases.

EXAMPLE:	*RESPONSE*	*"POLICE"*
	KEY	*"SECRET POLICE"*

2. Response string includes both acronym and expansion (where they appear juxtaposed in the text) instead of just one or the other.

EXAMPLE:	*RESPONSE*	*"ARMY OF NATIONAL LIBERATION (ELN)"*
	KEY	*"ARMY OF NATIONAL LIBERATION" / "ELN"*

PARTIALLY CORRECT:

The number of cases where it is justifiable to score this slot partially correct should be very limited.

3.2.7 Slot 7 -- PERPETRATOR: CONFIDENCE

FULLY CORRECT:

Mismatch not allowed to be scored fully correct.

PARTIALLY CORRECT:

The number of cases where it is justifiable to score this slot partially correct should be limited, especially for situations other than the following: System determines a a lesser confidence than actually warranted: POSSIBLE (system response) instead of CLAIMED OR ADMITTED, SUSPECTED OR ACCUSED, or SUSPECTED OR ACCUSED BY AUTHORITIES (key). Even in these cases, there has to be some strong justification based on e.g. a difference of opinion as to how a human would interpret the text in order to justify partial correctness.

NOTE: The scoring program will automatically score the system response partially correct in the case where the system generates SUSPECTED OR ACCUSED instead of SUSPECTED OR ACCUSED BY AUTHORITIES.

3.2.8 Slot 8 -- PHYSICAL TARGET: ID(S)

FULLY CORRECT:

See section 3.1.1.1.

PARTIALLY CORRECT:

1. See section 3.1.1.2.

2. Response string is good enough to corroborate categorization made in TYPE slot (assuming system response for TYPE slot is correct). Note that the string in the key may sometimes not be good enough by this criterion; in such cases you must decide for yourself whether the system response is as good as he filler in the key is.

3.2.9 Slot 9 -- PHYSICAL TARGET: TOTAL NUM

PARTIALLY CORRECT:

System response is PLURAL instead of a specific number in the key, in cases where filler had to be summed up, especially where approximate numbers are given, e.g., "some 20 power stations and over 30 banks".

3.2.10 Slot 10 -- PHYSICAL TARGET: TYPE(S)

FULLY CORRECT:

Mismatch not allowed to be scored fully correct.

PARTIALLY CORRECT:

The number of cases where it is justifiable to score this slot partially correct should be very limited.

3.2.11 Slot 11 -- HUMAN TARGET: ID(S)

FULLY CORRECT:

1. See section 3.1.1.1.

2. Response is a correct proper name, but person's title/role is included as part of name, rather than in parentheses following the name.

EXAMPLE:	*RESPONSE*	*"MR. XYZ"*
	KEY	*"XYZ" ("MR.")*

PARTIALLY CORRECT:

1. See section 3.1.1.2.

2. Response is a correct proper name, but person's title/role is missing or incorrect.

EXAMPLE:	*RESPONSE*	*"XYZ"*
	KEY	*"XYZ" ("MR.")*

3.2.12 Slot 12 -- HUMAN TARGET: TOTAL NUM

PARTIALLY CORRECT:

System response is PLURAL instead of a specific number in the key, in cases where filler had to be summed up, especially where approximate numbers are given, e.g., "some 20 employees and over 30 other people".

3.2.13 Slot 13 -- HUMAN TARGET: TYPE(S)

FULLY CORRECT:

Mismatch not allowed to be scored fully correct.

PARTIALLY CORRECT:

The number of cases where it is justifiable to score this slot partially correct should be limited, especially for situations other than the following, where "partially correct" may be justified if the text is particularly unclear:

1. System response is GOVERNMENT OFFICIAL or ACTIVE MILITARY; key has FORMER GOVERNMENT OFFICIAL or FORMER ACTIVE MILITARY.

2. System response is POLITICAL FIGURE; key has GOVERNMENT OFFICIAL.

3.2.14 Slot 14 -- TARGET: FOREIGN NATION(S)

FULLY CORRECT:

Mismatch not allowed to be scored fully correct.

PARTIALLY CORRECT:

The number of cases where it is justifiable to score this slot partially correct should be limited, especially for situations other than the following: System responds with correct country, but in a form that doesn't match the set list.

EXAMPLE:	*RESPONSE*	*U.S.*
	KEY	*UNITED STATES*

3.2.15 Slot 15 -- INSTRUMENT: TYPE(S)

FULLY CORRECT:

Mismatch not allowed to be scored fully correct.

PARTIALLY CORRECT:

The number of cases where it is justifiable to score this slot partially correct should be extremely limited, except in those cases that are handled automatically by the scoring program, i.e., where the system response is a set list item that is a superset of the filler in the key, as determined by the shallow hierarchy of instrument types provided in the task documentation..

EXAMPLE:	*RESPONSE*	*GUN*
	KEY	*MACHINE GUN*

3.2.16 Slot 16 -- LOCATION OF INCIDENT

PARTIALLY CORRECT:

1. The key expresses a range between two known locations, and the system response contains only one location.

EXAMPLE:	*RESPONSE*	*COLOMBIA: MEDELLIN (CITY)*
	KEY	*COLOMBIA: MEDELLIN (CITY) - CALI (CITY)*

2. Response has correct country, but in a form that doesn't match the set list.

EXAMPLE:	*RESPONSE*	*U.S.*
	KEY	*UNITED STATES*

NOTE: The scoring program will automatically score a response partially correct when it contains correct country but no specific place or an incorrect specific place.

EXAMPLES:	*RESPONSE*	*COLOMBIA*
	KEY	*COLOMBIA: MEDELLIN (CITY)*

	RESPONSE	*COLOMBIA: CALI (CITY)*
	KEY	*COLOMBIA: ANTIOQUIA (DEPARTMENT): MEDELLIN (CITY)*

3.2.17 Slot 17 -- EFFECT ON PHYSICAL TARGET(S)

FULLY CORRECT:

Mismatch not allowed to be scored fully correct.

PARTIALLY CORRECT:

The number of cases where it is justifiable to score this slot partially correct should be limited, especially for situations other than the following: System response correctly indicates that damage was done but under- or overestimates amount of damage.

EXAMPLE:	*RESPONSE*	*SOME DAMAGE*
	KEY	*DESTROYED*

3.2.18 Slot 18 -- EFFECT ON HUMAN TARGET(S)

FULLY CORRECT:

Mismatch not allowed to be scored fully correct.

PARTIALLY CORRECT:

The number of cases where it is justifiable to score this slot partially correct should be limited, especially for situations other than the following: System response contains less information than the key.

EXAMPLE:	*RESPONSE*	*NO INJURY*
	KEY	*NO INJURY OR DEATH*