APPENDIX C:

GUIDELINES FOR SCORING MISMATCHES BETWEEN SYSTEM RESPONSES AND ANSWER KEY

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. 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 NRaD along with your results. If there is sufficient evidence that errors were made that affect the scores obtained, a new key wil 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 to learn how to save your history buffer to a file for use in other scoring sessions required for completing the test procedure.

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

In the case of a mismatch on fillers for string-fill slots, the scoring program will permit you to score the response as fully correct, partially correct, or incorrect.

3.1.1.1 Fully Correct

NRaD 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. It is reasonable, for example, to assign full credit if your system picks up a string that is equivalent in meaning to the one in the key (e.g., "urban guerrillas" vs. "urban terrorists" in the PERP: INDIVIDUAL ID slot) but comes from a portion of the text that is distant from the portion containing most of the slot-filler 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). The scoring program attempts to do this automatically, but it does not have an exhaustive list of nonessential modifiers.

EXAMPLE (slot 19): *RESPONSE* *"THE 3 PEASANTS"*
KEY *"PEASANTS"*

In filling the key templates, such nonessential modifiers were generally included in the individual perpetrator ID slot (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 19): *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 19): *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 18): *RESPONSE* *"TORRES"*
KEY *"ALBERTO ROBERTO TORRES"*

(slot 10): *RESPONSE* *"BRIGADE"*
KEY *"6TH INFANTRY BRIGADE"*

Finally, scoring a filler in the INSTRUMENT: ID, PHYS TGT: ID, HUM TGT: NAME, or HUM TGT: DESCRIPTION slot partially correct is appropriate if the response string is not as good as the key but is good enough to corroborate categorization made in the corresponding TYPE slot, assuming system response for TYPE slot is correct.

EXAMPLE (slots 12 and 13): *RESPONSE* *"OIL"*
ENERGY: "OIL"
KEY *"OIL PIPELINE" / "PIPELINE"*
ENERGY: "OIL PIPELINE" / "PIPELINE"

3.1.1.3 "Distributed" Partially Correct

As described in section 5.2 of the MUC Scoring System User's Manual, the scoring program allows the user to "distribute" a partially correct score for a response across multiple key values. This action causes the scoring program to give the system credit for multiple partially correct fillers even though it only generated one. This is not allowed for set-fill slots, which are scored fully automatically, but it is allowed for other types of slots. The user is likely to find occasion to make use of this functionality primarily when scoring the target id/description/number slots.

EXAMPLE (slot 12): *RESPONSE* *"VEHICLES"*
KEY *"AMBULANCE"*
"FUEL TRUCK"
"STATION WAGON"

3.1.2 Set Fills

In the case of a mismatch on fillers for set-fill slots, the scoring program normally will automatically count the filler incorrect. But under certain conditions it will automatically assign partial credit instead (see subsections of section 3.2).

Set-fill slots that include cross-reference tags are scored automatically as follows:

SET-FILL VALUE	+	CROSS-REFERENCE TAG	=	SLOT SCORE
correct		correct		correct
correct		not correct		partial
partial		any		partial
incorrect		any		incorrect
missing		any		missing
spurious		any		spurious

NOTE: The LOCATION slot is not treated by the scoring program as having set fills.

3.1.3 Other Types of Fills

In the case of a mismatch on fillers for slots requiring other types of fills, the scoring program will normally query you to score the fillers as fully correct, partially correct, or incorrect. (However, assignment of partial credit for the LOCATION slot is sometimes assigned automatically -- see section 3.2.3.) Section 3.1.1.3, above, describes "distributed" partially correct score assignment.

The only non-set-fill slots that include cross-reference tags are HUM TGT: DESCRIPTION, HUM TGT: NUMBER, and PHYS TGT: NUMBER. Notes on scoring these slots are found in the appropriate subsections of section 3.2.

NRaD 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.

3.2 By Individual Slot

3.2.1 Slot 1 -- MESSAGE: TEMPLATE

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 the section below on INCIDENT: TYPE.) If there is no match on the incident type, manually mapping to NIL is allowed, at the discretion of the user.

If the user wishes to make a template map to a different one than the one determined by the automatic mapping algorithm, the scoring program will permit it as long as the content-based mapping conditions are met. The content-based

mapping conditions require at least a partial match on INCIDENT: TYPE, plus at least a partial match on at least one of the perpetrator slots (INDIV ID or ORG ID), one of the physical target slots (ID or TYPE), or one of the human target slots (NAME, DESCRIPTION, or TYPE).

3.2.2 Slot 2 -- INCIDENT: DATE

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).

EXAMPLE:	*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 has the date of the message dateline as the upper anchor or as its simple value.

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

NOTE: The system response should be judged INCORRECT when it is a default-looking value (as described above) in which the upper anchor does not match the key's simple date or its upper anchor.

EXAMPLE:	*RESPONSE*	*- 26 AUG 89 (default-looking fill)*
	KEY	*16 AUG 89*
	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 -- INCIDENT: LOCATION

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. The response is completely correct except for the country.

EXAMPLE: RESPONSE BOLIVIA: ANTIOQUIA (DEPARTMENT): MEDELLIN (CITY)
KEY COLOMBIA: ANTIOQUIA (DEPARTMENT): MEDELLIN (CITY)

NOTE: The scoring program will automatically score a response partially correct when it contains the correct country but no specific place. Partial credit can be interactively assigned when the response contains the correct country and an incorrect specific place.

EXAMPLE: RESPONSE COLOMBIA
KEY COLOMBIA: MEDELLIN (CITY)

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

RESPONSE COLOMBIA: CALI (CITY)
KEY COLOMBIA

3.2.4 Slot 4 -- INCIDENT: TYPE

The scoring system will automatically score all mismatches as incorrect, with the following exception: The scoring program will automatically score the slot partially correct in the case where the filler in the response is ATTACK and the filler in the key is any other incident type.

3.2.5 Slot 5 -- INCIDENT: STAGE OF EXECUTION

The scoring system will automatically score all mismatches as incorrect.

3.2.6 Slot 6 -- INCIDENT: INSTRUMENT ID

FULLY CORRECT: See section 3.1.1.1.

PARTIALLY CORRECT: See sections 3.1.1.2 and 3.1.1.3.

3.2.7 Slot 7 -- INCIDENT: INSTRUMENT TYPE

The scoring program will automatically score mismatching set fills incorrect, with the following exception: The scoring program will automatically score the fill partially correct when 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. This scoring is done irrespective of the correctness of the cross-reference tag.

EXAMPLE: *RESPONSE* *GUN: "AK-47"*
KEY *MACHINE GUN: "AK-47"*

RESPONSE *GUN: "BULLET"*
KEY *MACHINE GUN: "-"*

3.2.8 Slot 8 -- PERP: INCIDENT CATEGORY

The scoring system will automatically score all mismatches as incorrect.

3.2.9 Slot 9 -- PERP: INDIVIDUAL ID

FULLY CORRECT: See section 3.1.1.1.

PARTIALLY CORRECT:

1. See sections 3.1.1.2 and 3.1.1.3.

2. Key contains rather general data and the response contains consistent, but inferior, general strings.

EXAMPLE: *RESPONSE* *"TERRORIST ACTIONS"*
KEY *"URBAN TERRORISTS"*

3.2.10 Slot 10 -- PERP: ORGANIZATION ID

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: See sections 3.1.1.2 and 3.1.1.3.

3.2.11 Slot 11 -- PERP: ORGANIZATION CONFIDENCE

All mismatching set fills will automatically be scored incorrect, with the following exception: 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. This scoring is done irrespective of the correctness of the cross-reference tag.

3.2.12 Slot 12 -- PHYS TGT: ID

FULLY CORRECT: See section 3.1.1.1.

PARTIALLY CORRECT: See sections 3.1.1.2 and 3.1.1.3.

3.2.13 Slot 13 -- PHYS TGT: TYPE

The scoring program will automatically score mismatching set fills incorrect, with the following exception: The scoring program will automatically score the system response partially correct in the case where the system generates POLITICAL FIGURE OFFICE OR RESIDENCE instead of GOVERNMENT OFFICE OR RESIDENCE. This scoring is done irrespective of the correctness of the cross-reference tag.

3.2.14 Slot 14 -- PHYS TGT: NUMBER

PARTIALLY CORRECT:

The number of cases where it is justifiable to score this slot partially correct should be extremely limited, especially in cases other than the following: response has a single number, and key has a range which includes that number as an anchor; response has a single number, and key has a tilde in front of that same number. In such cases, partial credit may be assigned irrespective of the correctness of the cross-reference tag.

EXAMPLE:

RESPONSE	*7: "PYLONS"*	*or*	*7: "THINGS"*
KEY	*5 - 7: "PYLONS"*		*5 - 7: "PYLONS"*
RESPONSE	*7: "PYLONS"*	*or*	*7: "THINGS"*
KEY	*~7: "PYLONS"*		*5 - 7: "PYLONS"*

It is also possible to "distribute" a partially correct score across multiple key values, as described in section 3.1.1.3. It would be justifiable to do this only in those cases where distribution of a partially correct score had already been done on the referenced filler in the PHYS TGT: ID slot.

EXAMPLE:

RESPONSE	*3: "VEHICLES"*
KEY	*1: "AMBULANCE"*
	1: "FUEL TRUCK"
	1: "STATION WAGON"

3.2.15 Slot 15 -- PHYS TGT: FOREIGN NATION

The scoring program will automatically score mismatching set fills incorrect.

3.2.16 Slot 16 -- PHYS TGT: EFFECT OF INCIDENT

The scoring program will automatically score mismatching set fills incorrect, with the following exception: The scoring program will automatically score the fill partially correct if the system response is DESTROYED instead of SOME DAMAGE. (The reasoning here is that an understandable error would be to generate DESTROYED

rather than SOME DAMAGE if a text says that a bomb destroyed part of a target (e.g., a few offices in a building that is identified as a target) and doesn't explicitly say that this implies that the target as a whole was merely damaged.) This scoring is done irrespective of the correctness of the cross-reference tag.

3.2.17 Slot 17 -- PHYS TGT: TOTAL NUMBER

PARTIALLY CORRECT:

The number of cases where it is justifiable to score this slot partially correct should be extremely limited, especially in cases other than the following: response has a single number, and key has a range which includes that number as an anchor; response has a single number, and key has a tilde in front of that same number.

EXAMPLE: *RESPONSE* *7*
KEY *5 - 7*

RESPONSE *7*
KEY *~7*

3.2.18 Slot 18 -- HUM TGT: NAME

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 the HUM TGT: DESCRIPTION slot.

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

3.2.19 Slot 19 -- HUM TGT: DESCRIPTION

FULLY CORRECT: See section 3.1.1.1. However, when the filler for this slot includes a cross-reference tag, 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.

EXAMPLE: *RESPONSE* *"MAYOR": "TORRES"*
KEY *"MAYOR OF ACHI": "TORRES"*

PARTIALLY CORRECT:

1. See sections 3.1.1.2 and 3.1.1.3.

2. Filler has the correct title or role but includes the person's name.

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

3. The non-tag portion of the filler doesn't match the key but is deemed completely correct, and the cross-reference tag is incorrect or missing.

EXAMPLE: RESPONSE "MAYOR": "SANCHEZ"
KEY "MAYOR OF ACHI": "TORRES"

4. Scoring the entire filler partially correct may also be done if the non-tag portion of the filler is judged *partially* correct. 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: RESPONSE "FORMER MAYOR": "FULANO DE TAL"
KEY "SENATOR": "FULANO DE CUAL"
*(where "FORMER MAYOR" has been judged partially correct with respect to its *CORRECT* intended referent, "FULANO DE CUAL", i.e., on the basis of presuming that the whole system response was "FORMER MAYOR": "FULANO DE CUAL" rather than "FORMER MAYOR": "FULANO DE TAL")*

NOTE: 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.

3.2.20 Slot 20 -- HUM TGT: TYPE

The scoring program will automatically score mismatching set fills incorrect, with the exception of the following cases, where the scoring program will automatically score the filler partially correct:

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.

This scoring is done irrespective of the correctness of the cross-reference tag.

3.2.21 Slot 21 -- HUM TGT: NUMBER

PARTIALLY CORRECT:

The number of cases where it is justifiable to score this slot partially correct should be extremely limited, especially in cases other than the following: response has a single number, and key has a range which includes that number as an anchor; response has a single number, and key has a tilde in front of that same number. In such cases, partial credit may be assigned irrespective of the correctness of the cross-reference tag.

EXAMPLE: RESPONSE 7: "JUDGES" or 7: "PEOPLE"
KEY 5 - 7: "JUDGES" 5 - 7: "JUDGES"

RESPONSE 7: "JUDGES" or 7: "PEOPLE"
KEY ~7: "JUDGES" 5 - 7: "JUDGES"

It is also possible to "distribute" a partially correct score across multiple key values, as described in section 3.1.1.3. It would be justifiable to do this only in those cases

where distribution of a partially correct score had already been done on the referenced filler in the PHYS TGT: DESCRIPTION slot.

EXAMPLE: *RESPONSE* *3: "PEASANTS"*
KEY *1: "ADULT PEASANT"*
1: "TEEN-AGED PEASANT"
1: "BABY PEASANT"

3.2.22 Slot 22 -- HUM TGT: FOREIGN NATION

The scoring program will automatically score mismatching set fills incorrect.

3.2.23 Slot 23 -- HUM TGT: EFFECT OF INCIDENT

The scoring program will automatically score mismatching set fills incorrect, with the following exception: The scoring program will automatically score the fill partially correct if the response contains less information than the key.

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

RESPONSE *NO DEATH*
KEY *NO INJURY OR DEATH*

This scoring is done irrespective of the correctness of the cross-reference tag.

3.2.24 Slot 24 -- HUM TGT: TOTAL NUMBER

PARTIALLY CORRECT:

The number of cases where it is justifiable to score this slot partially correct should be extremely limited, especially in cases other than the following: response has a single number, and key has a range which includes that number as an anchor; response has a single number, and key has a tilde in front of that same number.

EXAMPLE: *RESPONSE* *7*
KEY *5 - 7*

RESPONSE *7*
KEY *~7*