

SRA: DESCRIPTION OF THE SOLOMON SYSTEM AS USED FOR MUC-4

Chinatsu Aone, Doug McKee, Sandy Shinn, Hatte Blejer

Systems Research and Applications (SRA)
2000 15th Street North
Arlington, VA 22201
aonec@sra.com

BACKGROUND

SRA's knowledge-based natural language processing system SOLOMON has been developed for text understanding since 1986. In addition to being a *domain-independent* NLP system, starting in the fall of 1990, SOLOMON has been extended as part of the MURASAKI project to become a *multi-lingual* text understanding system. It currently understands Spanish and Japanese as well as English texts. In order to achieve domain- and language-independence, SOLOMON separates data from processing modules. The processing modules do not assume any domain- or language-dependent facts; rather they are designed so that they work off separate data, i.e. lexicons, grammars, patterns, and knowledge bases, which vary according to the domain or language. To facilitate data acquisition, SRA has developed 2 tools: LEXTool for the development of lexicons and KBTool for the development of knowledge bases.

MUC-4 SYSTEM ARCHITECTURE

SRA's system as used for MUC-4 consists of the core NLP system SOLOMON, the Message Zoner, and Extract, as shown in Figure 1. SOLOMON consists of 5 processing modules: Preprocessing, Syntax, Semantics, Discourse and Pragmatics modules. The data SOLOMON used for MUC-4 consists of the lexicons, the grammar, the patterns, and the knowledge bases. In order to handle MUC-4 messages, the Message Zoner and the Pragmatics module were significantly extended, and the MUC-4 specific lexicons and knowledge bases were added to the existing data. In the following, each of the modules is explained along with examples from message TST2-MUC4-0048.

Message Zoner

The Message Zoner is the entry point for text into the MUC-4 Data Extraction system. It parses the free text areas of the incoming message into sections, tables, itemized lists, paragraphs, sentences, and individual tokens. This processing is domain-independent. The Zoner also parses the formatted header information for the particular message type. The Zoner's output is a canonical structure that we use for all of our projects, including projects which deal with non-English texts. Only paragraphs that contain certain MUC-specific keywords are processed by SOLOMON.

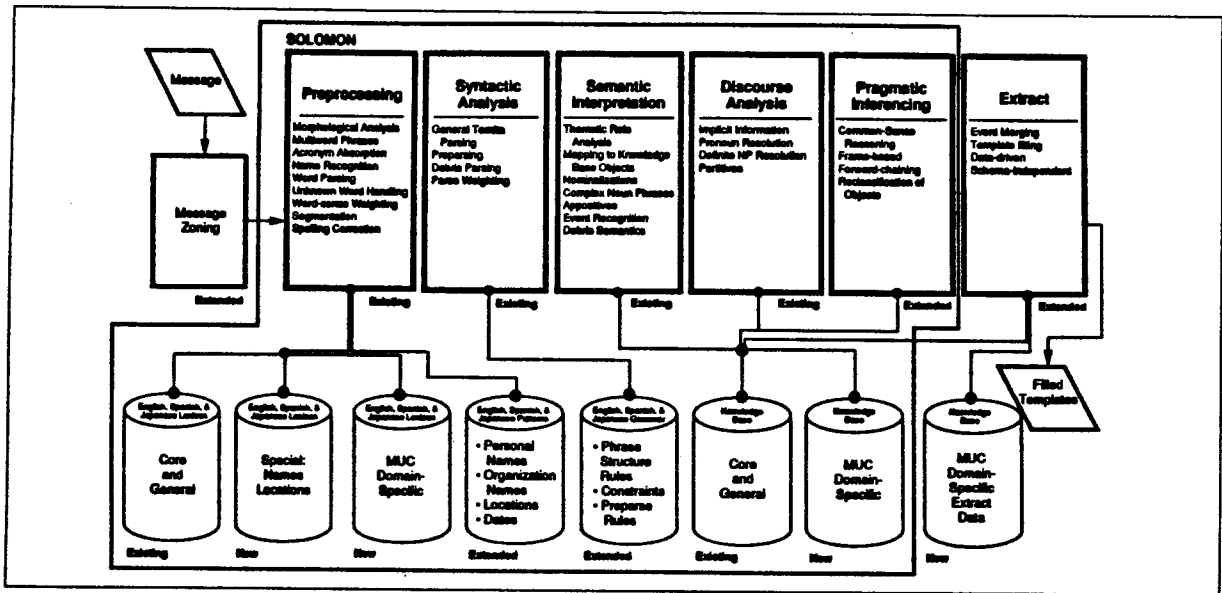


Figure 1: MUC-4 System Architecture

Preprocessing

The Preprocessing module performs word- and phrase-level analyses of input sentences. Since there are three types of lexicons, namely, the domain lexicons, the core lexicon and the “shallow” lexicon derived from a large corpus (i.e. the Dow Jones corpus from the Penn Treebank), when there is more than one entry with the same category for a word, the entry from the more specific lexicon is preferred.

In addition to regular lexical lookup and morphological analysis, the Preprocessing module uses various patterns to recognize productive multiwords and complex phrases like dates, personal names, organization names, locations, and so on. Also, it performs *acronym absorption*, where an acronym after a proper noun like “FARABUNDO MARTI NATIONAL LIBERATION FRONT (FMLN)” is removed from the output of preprocessing and learned by the system. The next time that acronym appears in isolation, preprocessing will understand that it has the same meaning as the original proper noun. Spelling correction and unknown word handling based on morphological endings are also performed.

Name Recognition

During preprocessing, proper names like “ALFREDO CRISTIANI” and “ROBERTO GARCIA ALVARADO” are dynamically recognized by the Spanish name pattern, which has been developed for the MURASAKI project, using the first names as anchors. The output of preprocessing for “ROBERTO GARCIA ALVARADO” is shown in Figure 2.

In addition, subsequent references to parts of these names, like “GARCIA”, are resolved using the information learned by the pattern. In this way, we do not need to put all the possible name combinations in the lexicon, but rather put only first names in the lexicon.

```
[ST: <PROPN>
ROOT: ROBERTO-GARCIA-ALVARADO
CATEGORY: PROP#
SEM-POINTER: MAN.310
NAME: (ROBERTO GARCIA ALVARADO)
INFLECTION: (3SG)
DEFINITE: T
ORIGIN: (SPANISH)
FIRST: ROBERTO
SURNAME: GARCIA]
```

Figure 2: Preprocessing: Name Pattern

Syntactic Analysis

Sentences are parsed using an X-bar-based phrase structure grammar, and SRA's custom modification of the Tomita parser, which handles Japanese and Spanish as well as English. The parser output is grammatical structures called Functionally Labelled Templates (FLT) which are built using a linguistic formalism that modifies and extends the f-structure of Lexical-Functional Grammar (LFG). These structures mark grammatical functions, like subject, object, specifier, and complement. Since the FLT formalism is language-independent, the same semantic interpretation module is used for all languages.

Preparsing

The MUC sentences are fairly long and complex, but in many cases SOLOMON will recognize major constituent boundaries using simple heuristics. For example, if a proper name is directly followed by a comma, some words, and another comma, then those words between the commas are assumed to be a constituent attaching to the proper name as an appositive (e.g. "ALFREDO CRISTIANI, NATIONALIST REPUBLICAN ALLIANCE (ARENA) PRESIDENT-ELECT,"). Other easily recognized probable constituents include "according to" phrases and "that" clauses following communication verbs. These smaller constituents are sent to general parsing in isolation before processing the entire sentence.

Debris Parsing

If general syntactic parsing of a sentence or constituent either fails or is taking too much time, the Debris Parsing module is invoked. First, the largest and best-weighted non-overlapping constituents recognized during parsing are extracted from the parse stack. The rest of the input is sent back into general parsing and debris parsing if necessary. When the entire sentence has been passed back to the parser, the resulting constituents are put together in a *debris* FLT. These structures are handled by a special submodule of Semantic Interpretation, called Debris Semantics.

Semantic Interpretation

The Semantic Interpretation module interprets the grammatical structures (FLT) to produce language-independent meaning representations called Semantically Labelled Templates (SLT). It performs semantic ambiguity resolution both during parsing (to reduce the number of parses) and during the construction of SLT (so that the best possible semantic interpretation is obtained.) The representation at this level is

```

Verb: 'accuse'
Situation Type: CAUSED-PROCESS
NLKB object for mapping: ACCUSE
Idiosyncracies: ((GOAL (MAPPING (LITERAL OF))))
Mapping:
(GOAL (MAPPING (LITERAL OF)) (TYPE SITUATION))
(AGENT (MAPPING (SURFACE SUBJECT)) (TYPE PERSON ORGANIZATION))
(THEME (MAPPING (SURFACE OBJECT)) (TYPE PERSON ORGANIZATION))

```

Figure 3: Mapping Information for “accuse”

language-independent because the representation language is based on the concepts in the knowledge bases which are shared among languages.

Verb mapping information is derived from both lexicons and KBs. In general, a lexical entry tells how each surface syntactic role is mapped to its corresponding thematic role, and a KB entry tells what the semantic type restrictions on these roles are. When necessary, however, lexical idiosyncracies, either syntactic or semantic, can be recorded in the lexicons. The mapping information for “accuse” is shown in Figure 3. The semantic concepts representing verbs like “accuse”, “condemn”, and “blame” are subclasses of a concept called JUDGEMENT-EVENT in our KB. The GOAL of this event (i.e. the embedded sentences under these verbs) are thus taken as facts, and mapped to the template as such.

Debris Semantics

When the Semantics module receives the output of Debris Parsing, it must process a collection of fragmentary syntactic constituents rather than a fully analyzed FLT. Debris Semantics will call general semantic interpretation on each of these constituents and fit them together as best it can based on semantic knowledge and constraints. This involves choosing a top-level S from the syntactic fragments, fitting the other fragments into it, and producing the most salient semantic interpretation for the sentence.

Nominalizations

Nominalized verbs, which often describe terrorist events as in “THE KILLING OF ATTORNEY GENERAL ROBERTO GARCIA ALVARADO”, “THE MURDER OF 10 UNION MEMBERS”, and “THE ATTACK ON FENASTRAS” are treated semantically like ordinary verbs. That is, the nouns “killing”, “murder”, and “attack” are mapped to event frames in the KBs (i.e. MURDER, KILL, and ATTACK respectively), and the modifying PPs of appropriate types become the THEME of these events, as in Figure 4.

Appositives

Both pre- and post-appositives like “ATTORNEY GENERAL ROBERTO GARCIA ALVARADO” and “MANUEL VALLEJO URIBE, A BUSINESSMAN” are interpreted so that the KB objects for the head nouns get additional class information provided by the appositives. In Figure 4, the appositive “ATTORNEY GENERAL” is interpreted so that the frame MAN.472 representing “ROBERT GARCIA ALVARADO” obtains additional ISA information (i.e. GOVERNMENT-OFFICIAL) from the appositive. This semantic interpretation enables resolution of the subsequent reference to the same man by “THE ATTORNEY GENERAL” in S21 (cf. Appendix A) in discourse processing.

```

(KILL.475 (ISA (VALUE KILL))
  (THEME (VALUE MAN.472))
  (UNIT (VALUE NATURAL-UNIT))
  (QUANTITY (VALUE (EXACT 1)))
  (SITUATION-TYPE (VALUE (CAUSED-PROCESS)))
  (DEFINITE (VALUE T))
  (ACTION-LINK (VALUE KILL.475))
  (ACTION-RELATION (VALUE RESULT)))

(MAN.472 (ISA (VALUE GOVERNMENT-OFFICIAL MAN))
  (NAMES (VALUE (ROBERTO GARCIA ALVARADO)))
  (QUANTITY (VALUE (EXACT 1)))
  (UNIT (VALUE NATURAL-UNIT)))

```

Figure 4: Semantics of “THE KILLING OF ATTORNEY GENERAL ROBERTO GARCIA ALVARADO”

Discourse Analysis

The Discourse Analysis module performs pronoun and definite NP resolution. Although this module handles some interesting phenomena such as partitives and super-subclass reference, this module needs the most work, especially to be able to handle phenomena which occur in other languages like Spanish and Japanese. Limited event discourse in terms of causality reasoning is done by Pragmatic Inferencing. For example, if it is mentioned that there was some terrorist attack and subsequently 3 people were found dead, we infer that the terrorist attack was the cause of the death of 3 people. Thus, we merge these 2 events into one terrorist event. We are planning to expand and incorporate the event discourse component into the Discourse module.

Partitives

SOLOMON handles partitives well because many of the domains for which it has been used call for understanding complex quantity expressions. The partitives like “FOUR OF THE VICE PRESIDENT’S CHILDREN” and “ONE OF THEM”, are interpreted by semantics so that the head noun (e.g. “ONE”, “FOUR”) represents a *part* of the object represented by the NP in the “of” phrase. The NP in the “of” phrase of the partitive construction must be a definite NP. Thus, getting the correct interpretation for partitives always requires correct definite anaphora resolution. In Figure 5, “THEM” in S22 was correctly resolved to “TWO BODYGUARDS”, which is represented by SECURITY-GUARD.292 in the SET-PARENT slot of ENTITY.299 representing “ONE”.

Reference by Superclass Concepts

The discourse resolution of “THE CRIME” to “KILLING” in S1 is handled by resorting to the KB hierarchy. One of SOLOMON’s anaphora resolution strategies is to look for an antecedent whose concept is a subclass of the concept represented by the anaphora. For example, in “John has a pet iguana, and he loves this lizard.”, “this lizard” is resolved to “a pet iguana” because the concept IGUANA is a subclass of the concept “LIZARD” in the KB.

The nominalized event reference “THE CRIME” is resolved in the same way. As explained earlier, a nominalized verb like “killing” is mapped to an event concept, in this case KILL, in the KB. The noun “crime” is mapped to the concept ANTI-CREATION-EVENT, which has subclasses like MURDER, ATTACK, BOMB-EVENT, DESTROY, and so on. KILL is also a subclass of ANTI-CREATION-EVENT, and therefore “THE CRIME” is resolved to “KILLING”. In this way, the two events are merged and a single

```

(ENTITY.299 (ISA (VALUE ENTITY))
  (QUANTITY (VALUE (EXACT 1)))
  (UNIT (VALUE NATURAL-UNIT))
  (SET-PARENT (VALUE SECURITY-GUARD.292))
  (TOKENS (VALUE (ONE OF THEM))))

(SEcurity-GUARD.292 (ISA (VALUE SECURITY-GUARD))
  (QUANTITY (VALUE (EXACT 2)))
  (UNIT (VALUE NATURAL-UNIT))
  (TOKENS (VALUE (THEM)
    (TWO BODYGUARDS))))

(INJURE.300 (ISA (VALUE INJURE))
  (SITUATION-TYPE (VALUE CAUSED-PROCESS))
  (THEME (VALUE ENTITY.299))
  (TENSE (VALUE PAST)))

```

Figure 5: Semantics of "ONE OF THEM WAS INJURED"

```

(MAN.478 (ISA (VALUE GOVERNMENT-OFFICIAL MAN))
  (NAMES (VALUE (FRANCISCO MERINO)))
  (QUANTITY (VALUE (EXACT 1)))
  (TOKENS
    (VALUE (VICE PRESIDENT ELECT FRANCISCO MERINO)))
  (UNIT (VALUE NATURAL-UNIT)))

(CIVILIAN-RESIDENCE.480 (ISA (VALUE CIVILIAN-RESIDENCE))
  (OCCUPIED-BY (VALUE MAN.478))
  (QUANTITY (VALUE (EXACT 1)))
  (UNIT (VALUE NATURAL-UNIT))
  (TOKENS (VALUE (MERINO 'S HOME))))

```

Figure 6: Semantics of "MERINO'S HOME"

template is created from S1.

Pragmatic Inferencing

This module was exploited extensively for the MUC-4 task in order to perform reasoning needed to go from literal interpretation of messages in our semantic representation to the MUC-4 template representation. For example, in S11 of message 0048, "MERINO'S HOME" should be categorized as GOVERNMENT OFFICE OR RESIDENCE because Merino is a vice president-elect. However, the default semantic type of "HOME" is CIVILIAN RESIDENCE, as shown in Figure 6. From this representation to the actual template, one must infer that a residence occupied by a government official is a government residence.

We made extensive use of the forward chainer of SRA's knowledge representation language TURNKEY for this kind of reasoning. It should be made clear that none of the forward rules are specific to particular terrorist incidents. Rather, all the rules reflect our commonsense reasoning. The rule which deals with the type of inference needed for the Merino example is Rule-025 in Figure 7.

In order to handle S12, where it should be determined that people in Merino's home were also targets, we added, after the final testing, another rule Rule-064, which says that any person inside a physical target is a human target.

```

(defrule rule-025 ((?x) (?x))
:example ("several homes of important government officials were looted..")
:if (and (or (civilian-residence ?x)
             (facility ?x))
        (occupied-by ?x ?occ)
        (or (government-official ?occ)
            (agency ?occ)
            (committee ?occ)
            (government ?occ)
            (international-governmental-organization ?occ)))
:then (government-facility ?x))

(defrule rule-064 ((?attack ?person) (?attack))
:example ("John's home was attacked." " there were 3 children in the home.")
:if (and (lisp-eval (script-general-muc-nature ?attack))
        (theme ?attack ?pt)
        (physical-area ?pt)
        (person ?person)
        (location ?person ?pt))
:then (theme ?attack ?person))

```

Figure 7: Forward Chaining Rules

Extract

The Extract module translates the domain-relevant portions of our language-independent meaning representation into database records. We maintain a strong distinction between code and data, and in fact use the same code to output to several databases; including flat template-style and more object-oriented schemas.

Given a top-level event for each processed sentence in the text, Extract decides what subevents of those top-level events can be assumed true and therefore extracted from. For example, if killing is condemned, as in S1, then that killing is mapped to the database.

We employ a fairly simple event merging strategy. Eventually we hope to handle this in discourse. Two events are merged when they have the same stage-of-execution, their “types” are compatible (i.e. either identical or one is just an attack), and one of the following conditions is met:

1. Both events have the same target.
2. Either event has no target.
3. Either event is only reporting deaths, injuries, or victims.

Unfortunately, this strategy does not merge the events in S21-22 with the event described in S1 since both incidents already have human targets. Of these merged events, Extract filters out those events which should not be mapped according to the rather complicated description provided in the MUC-4 task documentation.

To do the actual template filling, we rely on Extract data made up of kb-object/slot to db-table/field mapping rules and conversion functions for the individual values. For example, our AGENT slot in an ATTACK event corresponds to the PERPETRATOR fields in the MUC template. Information from the free text of the message is combined with that in the header when the text is not explicit about the date or location of the incidents.

APPENDIX

Sentences from TST2-MUC4-0048

S1: SALVADORAN PRESIDENT-ELECT ALFREDO CRISTIANI CONDEMNED THE TERRORIST KILLING OF ATTORNEY GENERAL ROBERTO GARCIA ALVARADO AND ACCUSED THE FARABUNDO MARTI NATIONAL LIBERATION FRONT (FMLN) OF THE CRIME.

S11-13: GUERRILLAS ATTACKED MERINO'S HOME IN SAN SALVADOR 5 DAYS AGO WITH EXPLOSIVES. THERE WERE SEVEN CHILDREN, INCLUDING FOUR OF THE VICE PRESIDENT'S CHILDREN, IN THE HOME AT THE TIME. A 15-YEAR-OLD NIECE OF MERINO'S WAS INJURED.

S21-22: ACCORDING TO THE POLICE AND GARCIA ALVARADO'S DRIVER, WHO ESCAPED UNSCATHED, THE ATTORNEY GENERAL WAS TRAVELING WITH TWO BODYGUARDS. ONE OF THEM WAS INJURED.

Generated Templates for TST2-MUC4-0048

0. MESSAGE: ID	TST2-MUC4-0048
1. MESSAGE: TEMPLATE	1
2. INCIDENT: DATE	- 19 APR 89
3. INCIDENT: LOCATION	EL SALVADOR
4. INCIDENT: TYPE	ATTACK
5. INCIDENT: STAGE OF EXECUTION	ACCOMPLISHED
6. INCIDENT: INSTRUMENT ID	"BOMB"
7. INCIDENT: INSTRUMENT TYPE	BOMB: "BOMB"
8. PERP: INCIDENT CATEGORY	TERRORIST ACT
9. PERP: INDIVIDUAL ID	"NO GROUP"
10. PERP: ORGANIZATION ID	"THE FARABUNDO MARTI NATIONAL LIBERATION FRONT"
11. PERP: ORGANIZATION CONFIDENCE	SUSPECTED OR ACCUSED BY AUTHORITIES: "THE FARABUNDO MARTI NATIONAL LIBERATION FRONT"
12. PHYS TGT: ID	-
13. PHYS TGT: TYPE	-
14. PHYS TGT: NUMBER	-
15. PHYS TGT: FOREIGN NATION	-
16. PHYS TGT: EFFECT OF INCIDENT	-
17. PHYS TGT: TOTAL NUMBER	-
18. HUM TGT: NAME	"ROBERTO GARCIA ALVARADO"
19. HUM TGT: DESCRIPTION	"ATTORNEY GENERAL": "ROBERTO GARCIA ALVARADO"
20. HUM TGT: TYPE	GOVERNMENT OFFICIAL: "ROBERTO GARCIA ALVARADO"
21. HUM TGT: NUMBER	1: "ROBERTO GARCIA ALVARADO"
22. HUM TGT: FOREIGN NATION	-
23. HUM TGT: EFFECT OF INCIDENT	DEATH: "ROBERTO GARCIA ALVARADO"
24. HUM TGT: TOTAL NUMBER	-
0. MESSAGE: ID	TST2-MUC4-0048
1. MESSAGE: TEMPLATE	2
2. INCIDENT: DATE	- 19 APR 89
3. INCIDENT: LOCATION	EL SALVADOR
4. INCIDENT: TYPE	BOMBING
5. INCIDENT: STAGE OF EXECUTION	ACCOMPLISHED
6. INCIDENT: INSTRUMENT ID	"A BOMB"
7. INCIDENT: INSTRUMENT TYPE	BOMB: "A BOMB"
8. PERP: INCIDENT CATEGORY	TERRORIST ACT
9. PERP: INDIVIDUAL ID	"AN INDIVIDUAL"
10. PERP: ORGANIZATION ID	"THE FARABUNDO MARTI NATIONAL LIBERATION FRONT"
11. PERP: ORGANIZATION CONFIDENCE	SUSPECTED OR ACCUSED BY AUTHORITIES: "THE FARABUNDO MARTI NATIONAL LIBERATION FRONT"
12. PHYS TGT: ID	-
13. PHYS TGT: TYPE	-
14. PHYS TGT: NUMBER	-

15. PHYS TGT: FOREIGN NATION	-
16. PHYS TGT: EFFECT OF INCIDENT	-
17. PHYS TGT: TOTAL NUMBER	-
18. HUM TGT: NAME	-
19. HUM TGT: DESCRIPTION	-
20. HUM TGT: TYPE	-
21. HUM TGT: NUMBER	-
22. HUM TGT: FOREIGN NATION	-
23. HUM TGT: EFFECT OF INCIDENT	-
24. HUM TGT: TOTAL NUMBER	-
0. MESSAGE: ID	TST2-MUC4-0048
1. MESSAGE: TEMPLATE	3
2. INCIDENT: DATE	14 APR 89
3. INCIDENT: LOCATION	EL SALVADOR: SAN SALVADOR (CITY)
4. INCIDENT: TYPE	BOMBING
5. INCIDENT: STAGE OF EXECUTION	ACCOMPLISHED
6. INCIDENT: INSTRUMENT ID	"BOMB"
	"EXPLOSIVES"
7. INCIDENT: INSTRUMENT TYPE	BOMB: "BOMB"
	EXPLOSIVE: "EXPLOSIVES"
8. PERP: INCIDENT CATEGORY	TERRORIST ACT
9. PERP: INDIVIDUAL ID	"GUERRILLAS"
10. PERP: ORGANIZATION ID	"THE FARABUNDO MARTI NATIONAL LIBERATION FRONT"
11. PERP: ORGANIZATION CONFIDENCE	SUSPECTED OR ACCUSED BY AUTHORITIES: "THE FARABUNDO MARTI NATIONAL LIBERATION FRONT"
12. PHYS TGT: ID	"MERINO'S HOME"
13. PHYS TGT: TYPE	GOVERNMENT OFFICE OR RESIDENCE: "MERINO'S HOME"
14. PHYS TGT: NUMBER	1: "MERINO'S HOME"
15. PHYS TGT: FOREIGN NATION	-
16. PHYS TGT: EFFECT OF INCIDENT	-
17. PHYS TGT: TOTAL NUMBER	-
18. HUM TGT: NAME	-
19. HUM TGT: DESCRIPTION	"THE VICE PRESIDENT'S CHILDREN IN THE HOME"
20. HUM TGT: TYPE	CIVILIAN: "THE VICE PRESIDENT'S CHILDREN IN THE HOME"
21. HUM TGT: NUMBER	PLURAL: "THE VICE PRESIDENT'S CHILDREN IN THE HOME"
22. HUM TGT: FOREIGN NATION	-
23. HUM TGT: EFFECT OF INCIDENT	DEATH: "THE VICE PRESIDENT'S CHILDREN IN THE HOME"
24. HUM TGT: TOTAL NUMBER	-
0. MESSAGE: ID	TST2-MUC4-0048
1. MESSAGE: TEMPLATE	4
2. INCIDENT: DATE	- 19 APR 89
3. INCIDENT: LOCATION	EL SALVADOR
4. INCIDENT: TYPE	BOMBING
5. INCIDENT: STAGE OF EXECUTION	ACCOMPLISHED
6. INCIDENT: INSTRUMENT ID	-
7. INCIDENT: INSTRUMENT TYPE	EXPLOSIVE: "-"
8. PERP: INCIDENT CATEGORY	TERRORIST ACT
9. PERP: INDIVIDUAL ID	-
10. PERP: ORGANIZATION ID	"THE FARABUNDO MARTI NATIONAL LIBERATION FRONT"
11. PERP: ORGANIZATION CONFIDENCE	SUSPECTED OR ACCUSED BY AUTHORITIES: "THE FARABUNDO MARTI NATIONAL LIBERATION FRONT"
12. PHYS TGT: ID	-
13. PHYS TGT: TYPE	-
14. PHYS TGT: NUMBER	-
15. PHYS TGT: FOREIGN NATION	-
16. PHYS TGT: EFFECT OF INCIDENT	-
17. PHYS TGT: TOTAL NUMBER	-
18. HUM TGT: NAME	"FRANCISCO MERINO"
19. HUM TGT: DESCRIPTION	"PRESIDENT ELECT": "FRANCISCO MERINO"
20. HUM TGT: TYPE	GOVERNMENT OFFICIAL: "FRANCISCO MERINO"
21. HUM TGT: NUMBER	1: "FRANCISCO MERINO"
22. HUM TGT: FOREIGN NATION	-
23. HUM TGT: EFFECT OF INCIDENT	-
24. HUM TGT: TOTAL NUMBER	-