PRC Inc: DESCRIPTION OF THE PAKTUS SYSTEM USED FOR MUC-4

Bruce Loatman

PRC Inc.
Technology Division
1500 PRC Drive
McLean, VA 22102
loatman_bruce@po.gis.prc.com

BACKGROUND

The PRC Adaptive Knowledge-based Text Understanding System (PAKTUS) has been under development as an Independent Research and Development project at PRC since 1984. It includes a core English lexicon and grammar, a concept network, processes for applying these to lexical, syntactic, semantic, and discourse analysis, and tools that support the adaptation of the generic core to new domains, primarily by acquiring sublanguage and domain-specific lexicon and conceptual topic patterns of interest. The lexical, syntactic, and semantic analysis components were completed before MUC-4 and required little adaptation. The discourse analysis component is new and was completed in the course of applying the system to MUC-4, although it is generic. The overall system is described in [1]. The present description concentrates on discourse analysis.

APPROACH

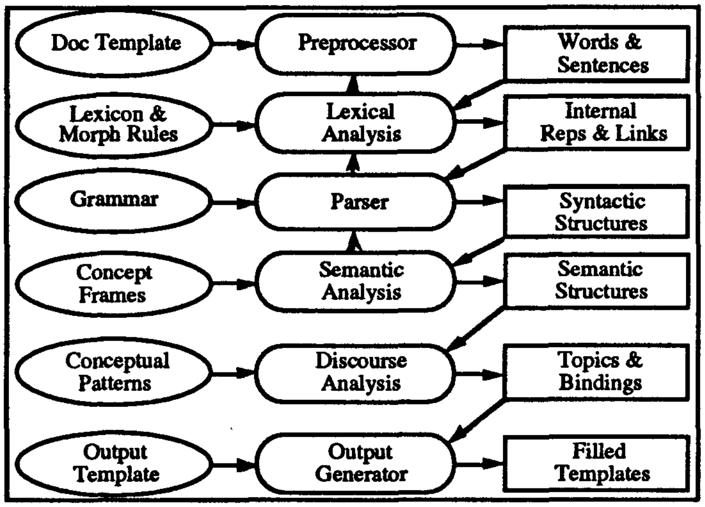


Figure 1. PAKTUS Architecture

The overall structure and operation of PAKTUS are shown in Figure 1. Processing proceeds mostly sequentially through preprocessing (the decomposition of the text stream into individual messages, message segments, sentences, and words), lexical analysis (morphological analysis and mapping of words into entries in the lexicon which contain information about their syntax and semantics), syntactic analysis (using a parser and grammar), semantic analysis (mapping the syntactic structures into conceptual frames with roles filled by phrase constituents), discourse analysis (identification of discourse topics and noun phrase reference), and finally extraction of information from discourse structures into domain-specific templates. The primary exception to sequential control flow is the interaction between the syntactic and semantic components at the clause and noun phrase level. This results in essentially deterministic parsing in linear time: the first syntactico-semantically successful parse of a sentence is accepted; others are never generated. Moreover, parse time is restricted, and the longest substring, along with any initial substring successfully parsed, is returned when parse time is exhausted.

Figure 2 shows the discourse analysis module, which was first used for MUC-4, and its interaction with the extraction module. The discourse module is generic for expository text, such as news reports. In figure 2, only the conceptual patterns and filter are MUC-4-specific, and these are part of the extraction component, not discourse analysis.

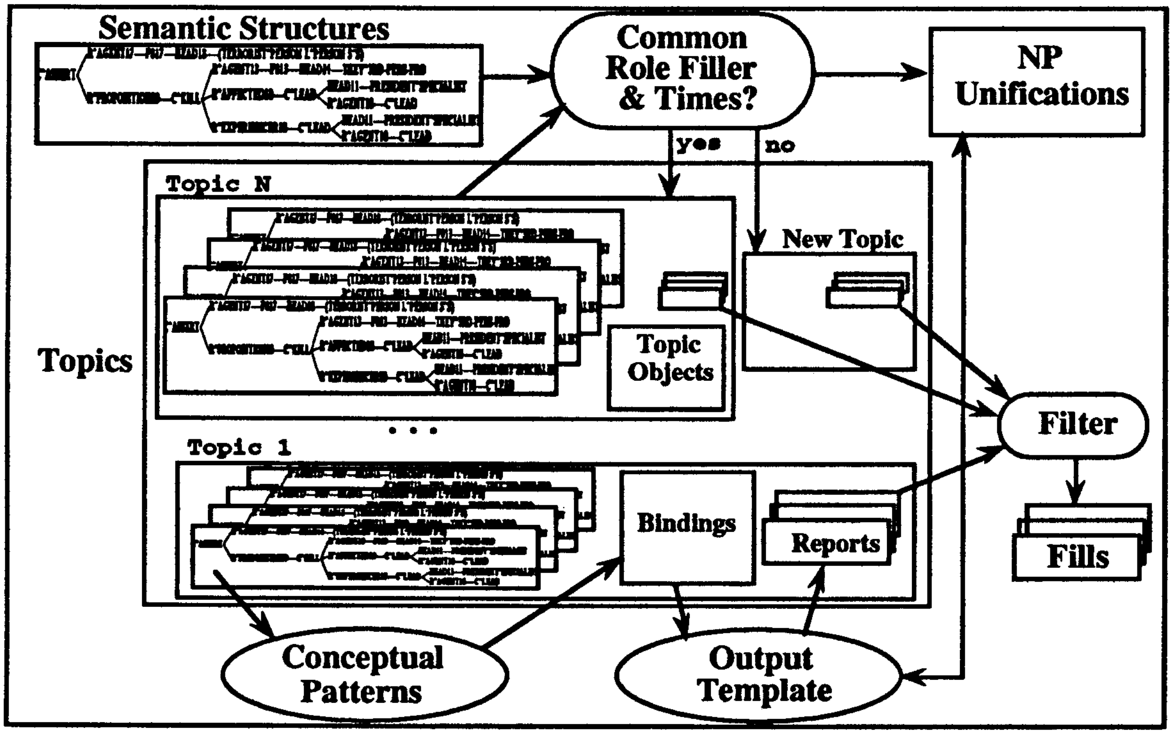


Figure 2. Discourse Analysis and Extraction Details

The discourse module operates on the semantic structures (case frames) produced by the semantic analysis module. It builds topic structures consisting of sets of case frames that have common topic objects and times. Topic objects are defined as fillers of certain case roles, specifically, 16 of the total 40 case roles used in PAKTUS, as illustrated in Figure 3. The most notable case role that is *excluded* as a topic object is the Agent. This is because topic structures are meant to represent information about entities that are being affected or focused upon in some way, whereas a single Agent can operate on several different entities. An example below, from the MUC-4 corpus, will clarify the importance of excluding the Agent as a topic role.

A side effect of comparing topic objects for commonality, is that some noun phrases (NPs) will be unified (i.e., considered by discourse analysis to have the same referent). It is possible (actually quite common) for two NPs to be considered common topic objects, but not be unified (e.g., in one MUC-4 passage, PAKTUS considers "crime" and "killing" to have topic commonality, but does not unify them since "crime" is more general).

After all case frames have been assigned to topic structures, domain-specific conceptual patterns are compared to the case frames, topic-by-topic, binding pattern variables to information that is extracted and put into event reports whose format is specified by a domain-specific template.

The NP unifications assist in this process, effectively consolidating information that may be widely dispersed in the text. Note that a single topic structure may contain information on multiple events. The final step in the extraction process is to filter and merge the event reports (e.g., for MUC-4, ignoring events that are too old, and merging events that can not be distinguished in time or location), and format the results to the output file.

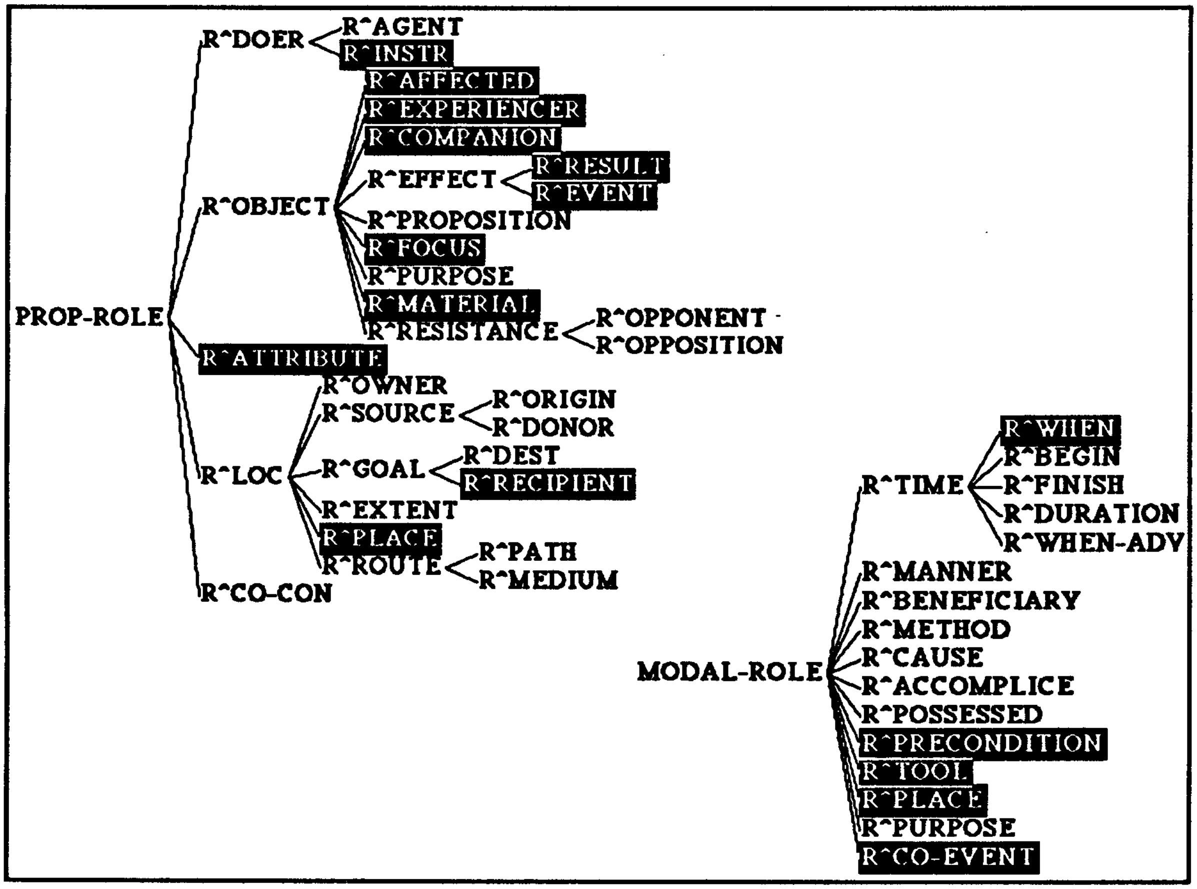
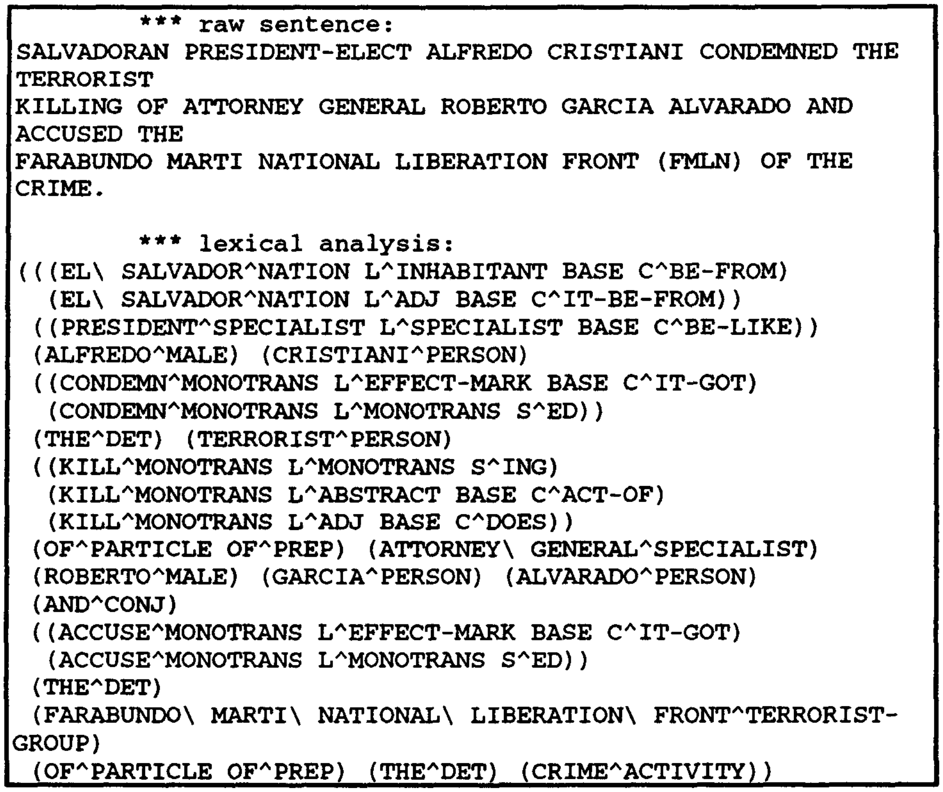


Figure 3. Case Roles Determining Topic Objects in PAKTUS

EXAMPLE OF MUC-4 DOCUMENT PROCESSING

Message number 48 from the "test2" set, which is reprinted in Appendix F, will be used to illustrate PAKTUS's operation for MUC-4. PAKTUS processes text sequentially, first stripping off the document header, then identifying sentences, which are processed syntactico-semantically one at a time, after which all the results are passed to the discourse component.

Figure 4 shows the raw, unprocessed text of the first sentence (S1), followed by its lexical analysis. Each word has one or more senses, represented as a root symbol, which is generally the concatenation of the English token, the "^" character, and the PAKTUS lexical category (e.g., "Condemn^Monotrans"), or as a simple structure involving a root, lexical category, inflectional mark, and sometimes a conceptual derivation (e.g. the structure "(Condemn^Monotrans L^Effect-mark Base C^It-got)" represents the adjective sense of "condemned"). For each word, all senses in the PAKTUS lexicon are fetched or derived at this time; disambiguation is generally delayed until the syntactic and semantic phases.



```
        *** raw sentence:
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.

        *** lexical analysis:
(((EL\ SALVADOR^NATION L^INHABITANT BASE C^BE-FROM)
  (EL\ SALVADOR^NATION L^ADJ BASE C^IT-BE-FROM))
 ((PRESIDENT^SPECIALIST L^SPECIALIST BASE C^BE-LIKE))
 (ALFREDO^MALE) (CRISTIANI^PERSON)
 ((CONDEMN^MONOTRANS L^EFFECT-MARK BASE C^IT-GOT)
  (CONDEMN^MONOTRANS L^MONOTRANS S^ED))
 (THE^DET) (TERRORIST^PERSON)
 ((KILL^MONOTRANS L^MONOTRANS S^ING)
  (KILL^MONOTRANS L^ABSTRACT BASE C^ACT-OF)
  (KILL^MONOTRANS L^ADJ BASE C^DOES))
 (OF^PARTICLE OF^PREP) (ATTORNEY\ GENERAL^SPECIALIST)
 (ROBERTO^MALE) (GARCIA^PERSON) (ALVARADO^PERSON)
 (AND^CONJ)
 ((ACCUSE^MONOTRANS L^EFFECT-MARK BASE C^IT-GOT)
  (ACCUSE^MONOTRANS L^MONOTRANS S^ED))
 (THE^DET)
 (FARABUNDO\ MARTI\ NATIONAL\ LIBERATION\ FRONT^TERRORIST-
GROUP)
 (OF^PARTICLE OF^PREP) (THE^DET) (CRIME^ACTIVITY))
```

Figure 4. Lexical Analysis of the First Sentence of Test2 Document Number 48

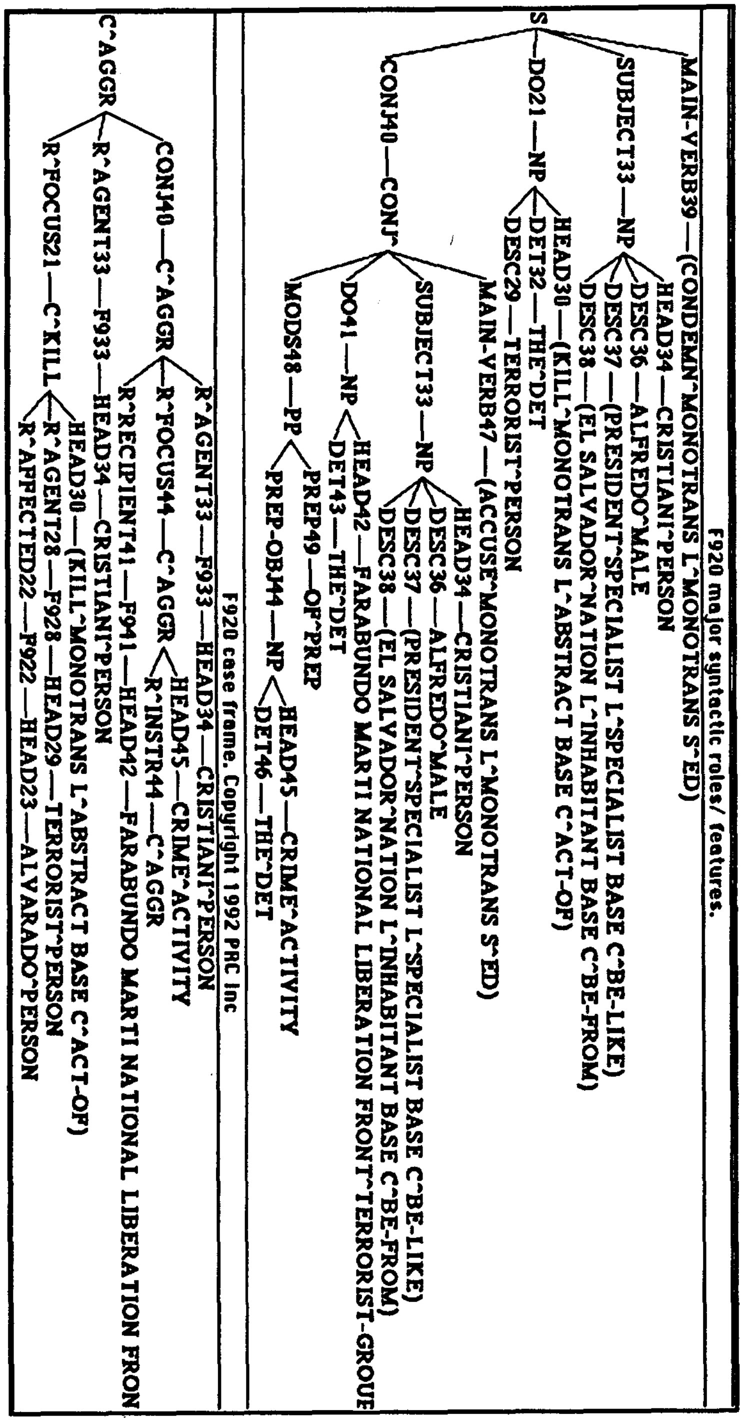
The syntactic and conceptual analyses of this sentence are shown in Figure 5. Note that conceptual structures are produced for some nouns (notably here, "killing"), not just for verbs. These conceptual structures are essential to the overall task of information extraction; if no case frames are produced for a sentence (i.e., the syntactico-semantic analysis failed), it is completely ignored by the discourse analysis and extraction processes.

The syntactic analysis produced for S1 is a configuration of syntactic registers (the main ones are shown in the figure) and register fillers. In this case, the main clause has a Main-Verb (condemned), Subject (the NP whose Head is Cristiani), and Direct Object (the terrorist killing NP). The conjoined clause ("and accused the FMLN ...") was correctly parsed, and its gap (no explicit Subject) has been filled in with the Subject of the main clause.

PAKTUS produced four case frames for S1, one for each of the two clauses, one for the "killing" NP and one for the "crime" NP.

This conceptual analysis will enable the discourse analysis module to determine that "the crime" refers to "the killing" because 1) both are topic objects (as fillers of Focus roles), 2) "crime's" concept C^AGGR is a generalization of "the killing's" concept of C^KILL in the PAKTUS conceptual network, and 3) "crime" appears in a subordinate clause (all three conditions are required for this reference resolution).

Determining that the crime refers to the killing here is important; it enables PAKTUS to identify the FMLN as the accused perpetrator of Alvarado's killing. It can also determine that the accusation is made by an authority (president-elect Cristiani), thanks to the gap filling by the syntactic analysis.



```
F920 major syntactic roles/ features.
  MAIN-VERB39—(CONDEMN^MONOTRANS L^MONOTRANS S^ED)
                    HEAD34—CRISTIANI^PERSON
  SUBJECT33—NP      DESC36—ALFREDO^MALE
                    DESC37—(PRESIDENT^SPECIALIST L^SPECIALIST BASE C^BE-LIKE)
                    DESC38—(EL SALVADOR^NATION L^INHABITANT BASE C^BE-FROM)
                HEAD30—(KILL^MONOTRANS L^ABSTRACT BASE C^ACT-OF)
S—DO21—NP       DET32—THE^DET
                DESC29—TERRORIST^PERSON
                MAIN-VERB47—(ACCUSE^MONOTRANS L^MONOTRANS S^ED)
                                  HEAD34—CRISTIANI^PERSON
                SUBJECT33—NP      DESC36—ALFREDO^MALE
                                  DESC37—(PRESIDENT^SPECIALIST L^SPECIALIST BASE C^BE-LIKE)
  CONJ40—CONJ^                    DESC38—(EL SALVADOR^NATION L^INHABITANT BASE C^BE-FROM)
                DO41—NP     HEAD42—FARABUNDO MARTI NATIONAL LIBERATION FRONT^TERRORIST-GROUE
                            DET43—THE^DET
                            PREP49—OF^PREP
                MODS48—PP                        HEAD45—CRIME^ACTIVITY
                            PREP-OBJ44—NP        DET46—THE^DET
F920 case frame. Copyright 1992 PRC Inc
                        R^AGENT33—F933—HEAD34—CRISTIANI^PERSON
       CONJ40—C^AGGR    R^FOCUS44—C^AGGR    HEAD45—CRIME^ACTIVITY
                                            R^INSTR44—C^AGGR
                        R^RECIPIENT41—F941—HEAD42—FARABUNDO MARTI NATIONAL LIBERATION FRON
C^AGGR R^AGENT33—F933—HEAD34—CRISTIANI^PERSON
                        HEAD30—(KILL^MONOTRANS L^ABSTRACT BASE C^ACT-OF)
       R^FOCUS21—C^KILL R^AGENT28—F928—HEAD29—TERRORIST^PERSON
                        R^AFFECTED22—F922—HEAD23—ALVARADO^PERSON
```

Figure 5. Syntactic and Conceptual Analysis of S1

The second sentence (S2) contains the phrase "Merino also declared that the death of the Attorney General..." which PAKTUS recognizes as referring to the killing in S1, so this phrase is consolidated into the same topic structure. Merino is not, however, a topic object, since he is in the Agent role, which is not a topic role. This is important, because later in the text there is a report of a guerrilla attack on Merino's home. Merino is a topic object there. That he is not a topic object in S2 enables PAKTUS to recognize this later passage as a different topic.

The complete filled templates for this article are shown in Figure 6 (the ordering of the templates is immaterial). They contain almost all of the information that should have been extracted. The only missing information is "no injury" to a bodyguard in template 2. Also, the city in which one incident (fill 2) occurred is incorrectly reported, and a "terrorist" perpetrator is reported redundantly in fill 2.

One item missing from template 1, compared to the official answer key, is the FMLN as the perpetrating organization. Nowhere in the text is it stated or implied that the FMLN attacked Merino's home, however. Much of the MUC-4-specific knowledge that would be encoded in conceptual patterns to identify information in topic structures to be extracted in the template fills, was not entered into PAKTUS, due to our development time and effort limits. The fills shown include some information (underlined in the figure) derived with minor enhancements that were made after the official test run for MUC-4, specifically, the addition of three conceptual patterns, correction of a minor error in the output specification of another one, and correcting a lexicon entry to mark "vehicle's" type as "transport vehicle."

```
0.  MESSAGE: ID                    TST2-MUC4-0048
1.  MESSAGE: TEMPLATE              1
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        "EXPLOSIVES"
7.  INCIDENT: INSTRUMENT TYPE      EXPLOSIVE: "EXPLOSIVES"
8.  PERP: INCIDENT CATEGORY        TERRORIST ACT
9.  PERP: INDIVIDUAL ID            "GUERRILLAS"
10. PERP: ORGANIZATION ID          -
11. PERP: ORGANIZATION CONFIDENCE  -
12. PHYS TGT: ID                   "MERINO'S HOME"
13. PHYS TGT: TYPE                 CIVILIAN 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           "VICE PRESIDENT'S CHILDREN"
                                   "SEVEN CHILDREN"
                                   "15-YEAR-OLD NIECE"
20. HUM TGT: TYPE                  CIVILIAN: "15-YEAR-OLD NIECE"
                                   CIVILIAN: "SEVEN CHILDREN"
                                   CIVILIAN: "VICE PRESIDENT'S CHILDREN"
21. HUM TGT: NUMBER                1: "15-YEAR-OLD NIECE"
                                   7: "SEVEN CHILDREN"
                                   4: "VICE PRESIDENT'S CHILDREN"
22. HUM TGT: FOREIGN NATION        -
23. HUM TGT: EFFECT OF INCIDENT    INJURY: "15-YEAR-OLD NIECE"
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: EL SALVADOR (CITY)
4.  INCIDENT: TYPE                 BOMBING
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            "SALVADORAN URBAN GUERRILLAS"
                                   "GUERRILLA"
10. PERP: ORGANIZATION ID          "FARABUNDO MARTI NATIONAL LIBERATION FRONT (FMLN)"
11. PERP: ORGANIZATION CONFIDENCE  SUSPECTED OR ACCUSED BY AUTHORITIES:  "FARABUNDO MARTI NATIONAL LIBERATION FRONT
(FMLN)"
12. PHYS TGT: ID                   "VEHICLE"
13. PHYS TGT: TYPE                 TRANSPORT VEHICLE: "VEHICLE"
14. PHYS TGT: NUMBER               1: "VEHICLE"
15. PHYS TGT: FOREIGN NATION       -
16. PHYS TGT: EFFECT OF INCIDENT   SOME DAMAGE: "VEHICLE"
17. PHYS TGT: TOTAL NUMBER         -
18. HUM TGT: NAME                  "ROBERTO GARCIA ALVARADO"
19. HUM TGT: DESCRIPTION           "ATTORNEY GENERAL": "ROBERTO GARCIA ALVARADO"
                                   "TWO BODYGUARDS"
                                   "GARCIA ALVARADO'S DRIVER"
20. HUM TGT: TYPE                  GOVERNMENT OFFICIAL: "ROBERTO GARCIA ALVARADO"
                                   SECURITY GUARD: "TWO BODYGUARDS"
                                   CIVILIAN: "GARCIA ALVARADO'S DRIVER"
21. HUM TGT: NUMBER                1: "ROBERTO GARCIA ALVARADO"
                                   2: "TWO BODYGUARDS"
                                   1: "GARCIA ALVARADO'S DRIVER"
22. HUM TGT: FOREIGN NATION        -
23. HUM TGT: EFFECT OF INCIDENT    DEATH: "ROBERTO GARCIA ALVARADO"
                                   NO INJURY: "GARCIA ALVARADO'S DRIVER"
24. HUM TGT: TOTAL NUMBER          -
```

Figure 6: Template Fills for Test2 Report 48

REFERENCE

[1] Loatman, B, "Description of the PAKTUS System Used for MUC-3", *Proceedings of the 3rd Message Understanding Conference*, San Mateo, CA: Morgan Kaufmann, 1991.