Improve Interoperability by Formalizing the Natural Language Parts of Military Messages

Dr. Matthias Hecking
FGAN/FKIE
Neuenahrer Straße 20
53343 Wachtberg-Werthhoven
GERMANY

hecking@fgan.de

ABSTRACT

Due to the new military operations other than war, the extent of the natural language parts in formalized military messages increases. In order to guarantee further automatic interoperability between different C4ISR systems three problems must be solved. First, the natural language text must be automatically transformed into a formal structure, which represents the meaning of the text. Second, the formal structure must be exchanged between the different C4ISR systems. And third, the representation of the meaning must be automatically inserted into the C4ISR database. In this paper, we show that Information Extraction techniques can be used as a possible solution to the first sub problem. As the formal structure for representing the meaning we use typed feature structures.

1.0 INTRODUCTION

Relevant military information is available in different forms. On one side of the spectrum, there are highly formalized military messages (e.g. cf.. [2]). On the other side, we can find natural language texts with little structure (e.g. documents, e-mails, web pages). Due to the new military deployments (military operations other than war) the *extent of the natural language parts in formalized military reports increases*. This means, that more and more relevant information is only available in natural language texts.

Interoperability between different C4ISR systems can be realized by replication (cf. [3]) or by sending formatted messages. In both cases, the production and processing of the exchanged information should be completely automatically. But, if the extent of the natural language parts increases, then a processing of the texts by humans is necessary before the information can be entered into the databases of the C4ISR systems. This is an interruption of the automatic interoperability process.

If the above-described problem of interruption should be solved, three sub problems must be handled. First, the natural language text must be *automatically transformed* into a *formal structure*, which represents the meaning of the text. Second, the formal structure must be *exchanged* between the different C4ISR systems. And third, the representation of the meaning must be automatically processed and the result *inserted* into the C4ISR database.

In this paper, we propose to use *Information Extraction* (IE) techniques to solve the first mentioned sub problem. IE is an engineering approach based on results of computational linguistics to build systems that process huge amount of texts of a specific sort. Each IE system is tailored to a specific domain and task. The result of the IE process (the meaning of the text) is represented in *feature-value structures*. These feature-value structures together with an ontology are the fundamental mechanism for representing the natural language meaning. The second sub problem can be solved easily. Instead of sending the natural

Paper presented at the RTO IST Symposium on "Coalition C4ISR Architectures and Information Exchange Capabilities", held in The Hague, The Netherlands, 27-28 September 2004, and published in RTO-MP-IST-042.

language text, the feature structure will be send. The third sub problem demands for a mapping between feature-value structures and relational databases. In our *SOKRATES project* we have developed a component, which realizes this mapping into the LC2IEDM database (cf. [4]).

This paper is structured as follows. First, we will introduce an extended view of the interoperability problem due to free-form texts. Then, information extraction for producing the formal representation of meaning (cf. [1], [14]) is shortly introduced. In the main part of the paper, we show how the natural language part of a military message is represented. The presented example is from our SOKRATES project (cf. [4], [5], [6], [7], [8], [9], [10], [11], [12], [16], [17]) in which IE techniques and feature structures are used to represent German battlefield reports.

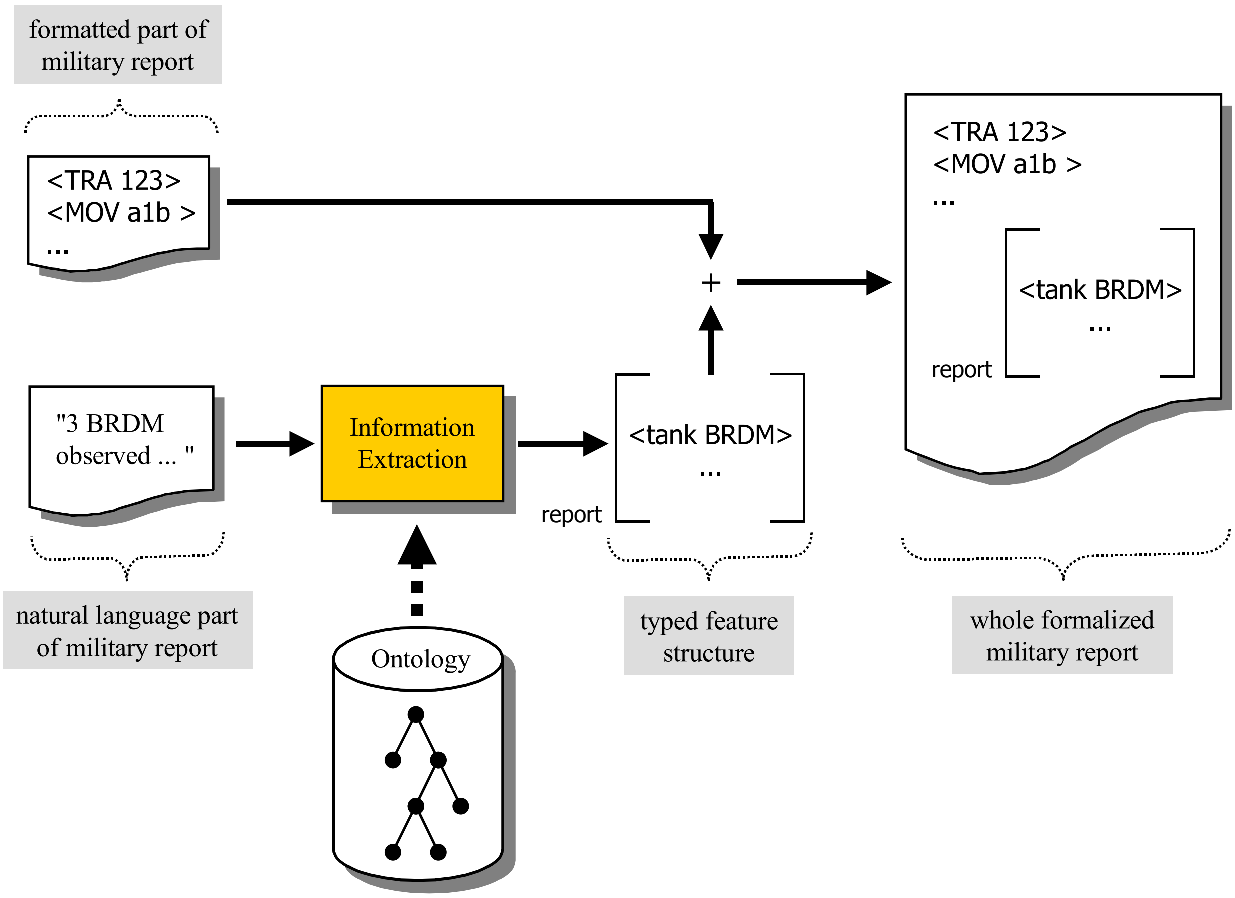

Figure 1: Combination of a formal and a natural language part in a military message

2.0 NATURAL LANGUAGE TEXT AND THE INTEROPERABILITY PROBLEM

The interoperability between different C4ISR systems assumes that the interchanged information is formatted according to a specific formal standard. In general, this standard is not designed to handle natural language text as part of military messages. Due to the new military operations other than war, the extent of the natural language parts in formatted military reports increases. Nevertheless, the automatic information exchange between the C4ISR systems should be still completely automatically. To realize this, the natural language text must be transformed into a formal representation of the meaning of the text. This semantic representation than can become a part of the formal standard.

In Figure 1, the situation is shown in more detail. On the left side, the formal part and the natural language part of a military report are shown. The written text ("3 BRDM observed …") must be transformed into a formal semantic representation. We propose to use Information Extraction (IE) technology. IE utilizes an ontology (a formal description of the domain concepts) to produce a formal description of the meaning (typed feature structure). Both parts of the report are now available in a formalized form and are combined. The whole formalized report is than used in the interoperability process.

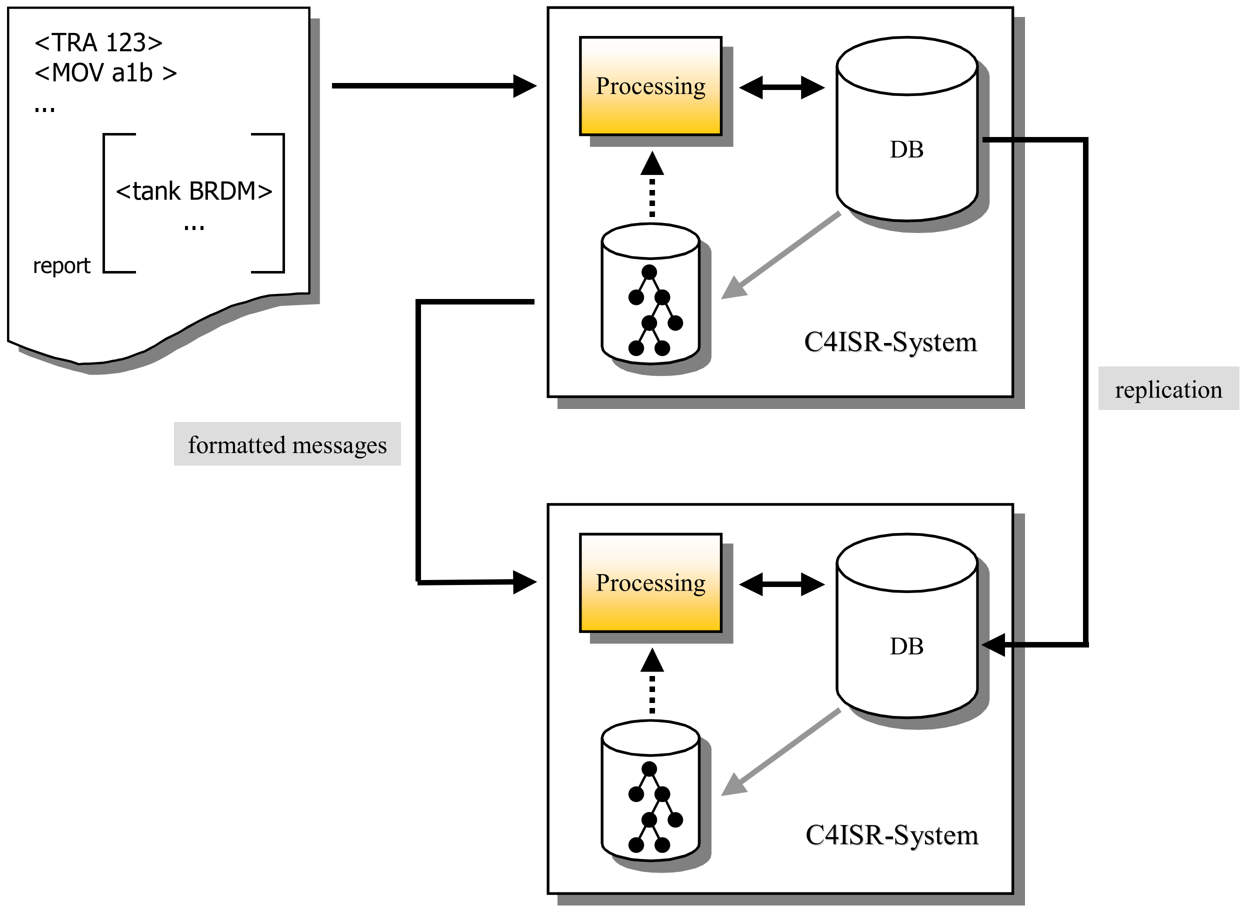

Figure 2: Processing of the formalized report in the C4ISR systems

In Figure 2 the processing of the whole formalized report is shown. This can be realized in two different ways. Reports can be interchanged directly by sending the produced formatted message. The other possibility is the replication of the database content. If the messages are send, in each C4ISR system the formal representation must be transformed into SQL statements. For those parts of the formal representation that origin from the natural language text, the processing must use the ontology from the IE process. This means, that the interoperability of the natural language parts of the messages create the new constraint that a common ontology must be used.

In the rest of this paper, one possibility of formalizing the meaning of military reports is presented. This approach was developed in the SOKRATES project. In this project, we applied IE to the analysis of German free-form battlefield reports. The SOKRATES prototype is able to process simple battlefield reports about movements (cf. [9], [10], [11], [12]). The objective of the SOKRATES project was not to show the applicability of the IE to the above described interoperability problem. But we believe that results from our project are a good initial point to find solutions for the mentioned interoperability problem.

3.0 INFORMATION EXTRACTION

The *information extraction* (IE) identifies, collects and normalizes relevant information about the Who, What, When, etc. from natural language texts (cf. [1], [14]). The relevant information is described through patterns called *templates*. These domain and task specific templates represent the meaning of the relevant information. During the IE task the templates are filled with the extracted information. Therefore, IE can be seen as the process of normalizing free-form text into a defined semantic structure.

To realize an IE system, language resources (lexicon, grammar) and appropriated parsing software are necessary. This software must be language-specific. Thus, the IE tools for the English language are not appropriated for analyzing German texts (due to the free-order of the language).

In order to achieve robust and efficient IE systems, domain knowledge (ontology) must be integrated and shallow algorithms must be used. The domain knowledge is tightly integrated with the language knowledge, e.g., the name 'Leopard' in the lexicon has the categorical information 'tank'. This association between words and semantic information is domain-specific and has to be change for other applications.

The IE process itself is divided into sub steps. After tokenizing the text, the sentence boundaries must be identified. Then, the morphological component identifies the word stems, the abbreviation, and detects the syntactic information (e.g., grammar case and gender). After this, the chunk parsing with transducers selects parts of the natural language text that are relevant for the specific information extraction task. The chunks are then used to instantiate the templates, which represent the result of the IE process.

```
(
   (:TYPE . :MILITARY-ITEM)
   (:LOCATION
      (
         (:COORDINATES . "32upd0290080100")
         (:TYPE . :POINT)
      )
   )
   (:QUALIFIER . :AT)
)
```

Figure 3: Example of a feature structure

The instantiated templates are a *formal description of the meaning* of the texts. In our research project SOKRATES and in this paper we use typed feature structures as templates. For more details about information extraction from battlefield reports refer to [10] and [12].

4.0 FORMALIZING THE NATURAL LANGUAGE PARTS

In this section we describe in detail, how the meaning of the natural language part of the military reports can be formalized. The meaning is represented by *typed feature structures*. The general feature structure formalism is widely used in computational linguistics, e.g., in the HPSG approach (cf. [15]). Each feature structure consists of an unordered quantity of *features* and a *type declaration*. In Figure 3 the type declaration is given by `(:TYPE . :MILITARY-ITEM)`. All other bracketed terms within the highest pair of brackets represent the set of the features. Each feature consists of a name (e.g., `:QUALIFIER`) and a value (e.g., `:AT`). In the simplest case the value is a number, a character string or an atomic type. In addition, a feature value can recursively be a feature structure, e.g., the value of the feature `:LOCATION` is a feature structure of the type `:POINT`.

If the IE processes the natural language part of a military report, the feature structures result from two sources. They can be part of a lexicon entry or they are constructed during the IE process.

During the development of an IE system an important development step is the determination and definition of the used feature structures. Because these structures can be based on each other, they are arranged in a hierarchy. In Figure 4, a part of the feature structure hierarchy is shown that is used in the SOKRATES IE module (the whole hierarchy can be found in [12]). The topmost feature structure is of the type **feature-structure**. This structure has no features. A possible subtype of **feature-structure** is the feature type **object** ("A ➔ B" means "A is a supertype of B."). This type has also various subtypes; one of them is the type **equipment**. **equipment** includes the features that are listed in square brackets under the type name. For each feature the name (e.g., hostility) and the feature type of the possible feature values (e.g., *hostility-type*) are given. The features are inherited down the hierarchy. So the types **vehicle**, **weapon** and also **antitank** have the same features as **equipment**. Some of the types are atomic and have no features, e.g., all subtypes of **nationality-code**. If features are defined for a subtype, these features extend the set of the inherited features. If the type of the feature value is enclosed by "*{...}*", the feature value consists of a set of feature structures of the indicated type. If the type is enclosed by "< ... >" the value consists of an ordered list of feature structures.

feature-structure
…
➔ **object**
➔ **equipment**
[
count *number*
hostility *hostility-type*
nationality *nationality-code*
abbreviation *string*
name *string*
qualifiers *{property-value}*
]
➔ **vehicle**
➔ **weapon**
➔ **antitank**
…

Figure 4: Part of the feature structure hierarchy

The meaning of the natural language part of a military report is represented by a feature structure that is one of the subtypes of the type **action**. The action type **move** (see Figure 5) is used to describe movements on the battlefield. The basic structure of **move** was adopted from the FrameNet project (cf. [13]) and adapted to the military requirements. During a movement, objects are moved. These objects are represented as feature values of feature "theme". The objects are moving from a starting point (source *military-item*) to an end point (goal *military-item*) using a path (path <*military-item*>). The objects can also be in an area (area *{military-item}*).

move
[
qualifiers *{property}*
area *{military-item}*
distance *number*
duration *number*
goal *military-item*
path <*military-item*>
source *military-item*
speed *number*
start-time *time*
theme *theme*
]

Figure 5: Feature structure 'move'

The feature "qualifiers" contains descriptions of adverbial phrases of the movement (e.g., "it moves *fast*"), if the descriptions are not contained in other features (e.g., in "goal <*military-item*>").

```
feature-structure
        …
    ➔       object
                ➔ equipment
                  [
                  count number
                  …
                  qualifiers {property-value}
                  ]
                        ➔ vehicle
                        ➔ weapon
                        …
                ➔ military-item
                  [
                  …
                  ]
                        ➔ facility
                                ➔ airfield
                                ➔ barrier
                        …
                ➔ partner
                        ➔ multiple-partner
                          [
                          first-partner single-partner
                          second-partner partner
                          ]
                        ➔ single-partner
                        …
```

Figure 6: Subtypes of type 'object'

In some of the feature values of the type **move**, structures of the subtypes of **object** are used. In Figure 6 these subtypes are shown. Ordnance can be found under **equipment**. Other military objects and concepts (e.g., airfield, barrier) are subtypes of **military-item**.

In order to describe movements, concepts of locations and directions are necessary. Directions are formalized by the type **direction** (see Figure 7), locations by **location** (see Figure 8) and its subtypes.

feature-structure
…
➔ **direction**
[
directed-to *dir-code*
qualifier-to *dir-qualifier*
]
➔ **dir-code**
➔ **east**
➔ **north**
➔ **northeast**
…

Figure 7: Formalization of directions

feature-structure
…
➔ **location**
[
name *string*
]
…
➔ **point**
[
coordinates *string*
]
…

Figure 8: Formalization of locations

In the following, we show how the meaning of the natural language part of a military battlefield report is represented with the introduced feature structures. The part is:

"2 BRDM 2 und 1 PT 76 durchfahren in rascher Fahrt auf Straße 14 – 1
die Ortschaft SULJIN HAN (CQ 5458) nach Westen."

("2 BRDM 2 and 1 PT 76 pass through the locality SULJIN HAN (CQ 5458) on the road 14 – 1 in fast travel to the west.")

In the tables of Figure 9, Figure 10, Figure 11 and Figure 12 the different feature structures are shown, which are produced by the IE from various parts of the text. In these tables is specified,

- which text part induces
- which syntactical phrase and
- the constructed feature structure (Feature structure A).

Also shown is:

- the syntactical function and
- the feature structure that will be part of the **move**-structure (Feature structure B).

<table>
<tr><td>Report part</td><td>2 BRDM 2</td><td>und</td><td>1 PT 76</td></tr>
<tr><td rowspan="2">Phrases</td><td>NP_1</td><td>Conj</td><td>NP_2</td></tr>
<tr><td colspan="3">NP</td></tr>
<tr><td>Feature structure A</td><td><pre>(
  (:TYPE.:TANK)
  (:ABB."brdm-2")
  (:COUNT . 2)
)</pre></td><td><pre>:SET</pre></td><td><pre>(
  (:TYPE . :TANK)
  (:ABB . "pt76")
  (:COUNT . 1)
)</pre></td></tr>
<tr><td>Syntactical function</td><td colspan="3">Subject</td></tr>
<tr><td>Feature structure B</td><td colspan="3"><pre>(:OBJECTS
  (:SET
    (
      (
        (:TYPE . :TANK)
        (:ABBREVIATION . "brdm-2")
        (:COUNT . 2)
)
      (
        (:TYPE . :TANK)
        (:ABBREVIATION . "pt76")
        (:COUNT . 1)
)
)
)
)</pre></td></tr>
</table>

Figure 9: Example – I

The part *"2 BRDM 2 und 1 PT 76"* is a nominal phrase (NP), which is constructed from two smaller phrases and a conjunction. For each of these phrases a feature structure of the type `:TANK` is produced. The whole NP is the subject of the sentence. Both **tank** feature structures are stored as values of the *objects* feature in the **move** structure.

Report part	***durchfahren***	***in rascher Fahrt***
Phrases	VP	PP

Feature structure A	`(:TYPE . :MOVE)`	`(`
		`(:VALUE . :FAST)`
		`(:TYPE . :PROPERTY)`
		`)`

Syntactical function	Predicate	Adverbial phrase

Feature structure B	`(:TYPE . :MOVE)`	`(:QUALIFIERS`
		`(:SET`
		`(`
		`(`
		`(:VALUE . :FAST)`
		`(:TYPE . :PROPERTY)`
		`)`
		`)`
		`)`
		`)`

Figure 10: Example – II

The verb *"durchfahren"* ("to pass through") forms the verbal phrase (VP). This VP is the predicate of the sentence. It is the feature type `(:TYPE . :MOVE)` of the VP, which is responsible for selecting the correct **action** subtype, in this example the **move** feature structure.

The prepositional phrase (PP) *"in rascher Fahrt"* ("in fast travel") produces a feature structure of the type `:PROPERTY`. This PP is an adverbial phrase, which specifies in more detail the VP (the movement is fast, not slow). The feature structure of this PP is stored in the *qualifiers* feature of the **move** structure.

<table>
<tr><td>Report part</td><td>auf Straße 14 - 1</td><td>die Ortschaft SULJIN HAN (CQ 5458)</td></tr>
<tr><td>Phrases</td><td>PP</td><td>NP</td></tr>
<tr><td>Feature structure A</td><td><pre>(
  (:TYPE . :WAY)
  (:NAME . "14-1")
  (:QUALIFIER . :ON)
)</pre></td><td><pre>(
  (:LOCATION
    (
      (:COORDINATES . "cq5458")
      (:TYPE . :POINT)
)
)
  (:NAME "suljin han")
  (:TYPE . :TOWN)
)</pre></td></tr>
<tr><td>Syn. fct.</td><td>Adverbial phrase</td><td>Object</td></tr>
<tr><td>Feature structure B</td><td colspan="2"><pre>(:AREA
  (:SET
    (
      (
        (:TYPE . :WAY)
        (:NAME . "14-1")
        (:QUALIFIER . :ON)
)
      (
        (:LOCATION
          (
            (:COORDINATES . "cq5458")
            (:TYPE . :POINT)
)
)
        (:NAME "suljin han")
        (:TYPE . :TOWN)
)
)
)
)</pre></td></tr>
</table>

Figure 11: Example - III

The prepositional phrase *"auf Straße 14 - 1"* ("on the road 14 – 1") produces a feature structure of the type `:WAY`. It is also an adverbial phrase and it is stored in the *area* feature of the **move** structure.

The `:TOWN`-structure is also stored in the *area* feature. This structure results from the report part *"die Ortschaft SULJIN HAN (CQ 5458)"* ("the locality SULJIN HAN (CQ 5458)").

Report part	***nach Westen***	**.**
Phrases	PP	

Feature structure A	`(`
	`  (:QUALIFIER . :TOWARDS)`
	`  (:DIRECTED-TO . :WEST)`
	`  (:TYPE . :DIRECTION)`
	`)`	

Syn. fct.	Adverbial phrase	

Feature structure B	`(:GOAL`
	`  (`
	`    (:QUALIFIER . :TOWARDS)`
	`    (:DIRECTED-TO . :WEST)`
	`    (:TYPE . :DIRECTION)`
	`  )`
	`)`	

Figure 12: Example - IV

"nach Westen" ("to the west") produces a feature structure of the type `:DIRECTION`. This is stored in the *goal* feature.

The meaning of the whole natural language part of the report (all feature structures B) is represented as a value of type **move** of the `:MESSAGE`-feature (see Figure 13 and Figure 14). This is the result of the IE process, which is attached to the formatted part of the military report.

For more information concerning the feature structures and the information extraction refer to [12].

```
(:MESSAGE
  (:SET
      (
        (
          (:AREA
            (:SET
              (
                (
                  (:TYPE . :WAY)
                  (:NAME . "14-1")
                  (:QUALIFIER . :ON)
                )
              (
                  (:LOCATION
                    (
                      (:COORDINATES . "cq5458")
                      (:TYPE . :POINT)
                    )
                  )
                  (:NAME . "suljin han")
                  (:TYPE . :TOWN)
                )
              )
            )
          )
          (:QUALIFIERS
            (:SET
              (
                (
                  (:VALUE . :FAST)
                  (:TYPE . :PROPERTY)
                )
              )
            )
          )
          (:TYPE . :MOVE)
          (:GOAL
            (
              (:QUALIFIER . :TOWARDS)
              (:DIRECTED-TO . :WEST)
              (:TYPE . :DIRECTION)
            )
          )
```

Figure 13: The feature structure of the example - I

```
            (:THEME
              (
                (:OBJECTS
                  (:SET
                    (
                      (
                        (:TYPE . :TANK)
                        (:ABBREVIATION . "brdm-2")
                        (:COUNT . 2)
                      )
                      (
                        (:TYPE . :TANK)
                        (:ABBREVIATION . "pt76")
                        (:COUNT . 1)
                      )
                    )
                  )
                )
                (:TYPE . :THEME)
              )
            )
          )
        )
  )
)
```

Figure 14: The feature structure of the example - II

5.0 CONCLUSION

Due to the new military operations other than war, the extent of the natural language parts in formalized military reports increases. To assure further on the automatic interoperability between different C4ISR systems three problems must be solved. First, the natural language text must be automatically transformed into a formal structure, which represents the meaning of the text. Second, the formal structure must be exchanged between the different C4ISR systems. And third, the representation of the meaning must be automatically inserted into the C4ISR database.

In this paper, we first explained the extended view on the interoperability problem. Then, we introduced the Information Extraction technology. In the main part of the paper, we showed that information extraction techniques could be used to solve the first of the aforementioned sub problems. For this, we introduced typed feature structures, explained parts of the feature structure hierarchy and, we gave an extensive example how to formalize the natural language parts of a report.

REFERENCES

[1] Appelt, D. & Israel, D. *Introduction to Information Extraction Technology*. Stockholm: IJCAI-99 Tutorial, 1999, http://www.ai.sri.com/~appelt/ie-tutorial/.

[2] ATCCIS WP 5-7, Edition 5.0. *Overview of the Land C2 Information Exchange Data Model (LC2IEDM)*. NATO: ATCCIS PWG, 2002.

[3] ATCCIS WORKING PAPER 14-3. *ATCCIS REPLICATION MECHANISM (ARM) CONSOLIDATED SPECIFICATION*. ARM Baseline 2.0, 18 March 2002.

[4] Casals Elvira, X. *Translation from Semantically Enriched Linguistic Representations to SQL Statements for the Project SOKRATES*. FKIE-Bericht Nr. 76, Wachtberg: Forschungsgesellschaft für Angewandte Naturwissenschaft e.V., 2004.

[5] Casals Elvira, X. *Project SOKRATES: Processing of Headers for the Information Extraction Component*. FKIE-Bericht, Wachtberg: Forschungsgesellschaft für Angewandte Naturwissenschaft e.V., to appear, 2004.

[6] Frey, M. L., Schade, U. *Project SOKRATES: Modular Framework for Military Report Processing*. FKIE-Bericht Nr. 73, Wachtberg: Forschungsgesellschaft für Angewandte Naturwissenschaft e.V., 2004.

[7] Hecking, M. *Natural Language Access for C4I Systems*. In: RTO IST Symposium on Information Management Challenges in Achieving Coalition Interoperability, RTO MP-064, Quebec, Canada, 2001.

[8] Hecking, M. *Analysis of Spoken Input to C2 Systems*. In: "Proceedings of the 7th International Command and Control Research and Technology Symposium (ICCRTS)", Québec City, Kanada, 2002.

[9] Hecking, M. *Information Extraction from Battlefield Reports*. In: "Proceedings of the 8th International Command and Control Research and Technolgy Symposium (ICCRTS)", June 17-19, 2003, National Defense University, Washington, DC.

[10] Hecking, M. *Analysis of Free-form Battlefield Reports with Shallow Parsing Techniques*. In: Information Systems Technology Panel Symposium "Military Data and Information Fusion", October 20-22, 2003, Prague, Czech Republic.

[11] Hecking, M. *How to Represent the Content of Free-form Battlefield Reports*. In: "Proceedings of the 2004 Command and Control Research and Technolgy Symposium (CCRTS)", June 15-17, 2004, Loews Coronado Bay Resort, San Diego, California.

[12] Hecking, M. *Informationsextraktion aus militärischen Freitextmeldungen*. FKIE-Bericht Nr. 74, Wachtberg: Forschungsgesellschaft für Angewandte Naturwissenschaft e.V., 2004.

[13] FrameNet-Projekt. *Concept "Motion"*, http://www.icsi.berkeley.edu/~framenet/data/html/frames/Motion.html, 2003.

[14] Pazienza, M. T. (ed.) *Information Extraction*. Berlin, 1999.

[15] Pollard, C. & Sag, I. A. *Head-Driven Phrase Structure Grammar*. The University of Chicago Press, Chicago, London, 1994.

[16] Schade, U. *Ontologieentwicklung für Heeresanwendungen*. Forschungsgesellschaft für Angewandte Naturwissenschaften e.V. (FGAN), FKIE-Bericht Nr. 57 VS-NfD, 2003.

[17] Schade, U. *Towards an Ontology for Army Battle C2 Systems*. In: "Proceedings of the 8th International Command and Control Research and Technolgy Symposium (ICCRTS)", June 17-19, 2003, National Defense University, Washington, DC.