Multilingual Content Extraction Extended with Background Knowledge for Military Intelligence

Dr. Matthias Hecking
Fraunhofer FKIE
matthias.hecking@fkie.fraunhofer.de
Dr. Andreas Wotzlaw
University of Cologne
wotzlaw@informatik.uni-koeln.de
Ravi Coote
Fraunhofer FKIE
ravi.coote@fkie.fraunhofer.de

Outline

1. Introduction
2. Combined Deep and Shallow Parsing
3. Logical Inferences on Text Content
4. Background Knowledge
5. Conclusion, References

1. Introduction

Motivation/Problem description:

- Necessity to analyze the content of large quantities of intelligence reports and other documents written in different languages.
- During this information and knowledge exploration (content analysis) a formal description of the actions and involved entities is constructed.
- The extracted information can be combined and enhanced with background knowledge.
- Conclusions can be drawn from the extracted and enhanced information.
- Various approaches:
  - Shallow parsing, application specific combination of analysis results, used in current projects, Information Extraction, ZENON project.
  - Our mIE project.
  - …

1. Introduction - mIE Project – Main ideas

- Our approach: The project "Multilingual content analysis with semantic inference on military relevant texts" (mIE)
  - Combined deep and shallow parsing approach.
  - Extracted meaning of each sentence is formalized in formal logic .
  - Simple English and (very simple) Arabic texts can be processed.
  - The formalized content is extended with background knowledge (integration of WordNet and YAGO).
  - New conclusions (logical inferences) can be drawn; application of theorem provers and model builders.

1. Introduction - mIE Project – Logical inference I

The problem of drawing conclusions on texts and relevant background knowledge is formalized as a pair of a text and a hypothesis. The following is a typical example:

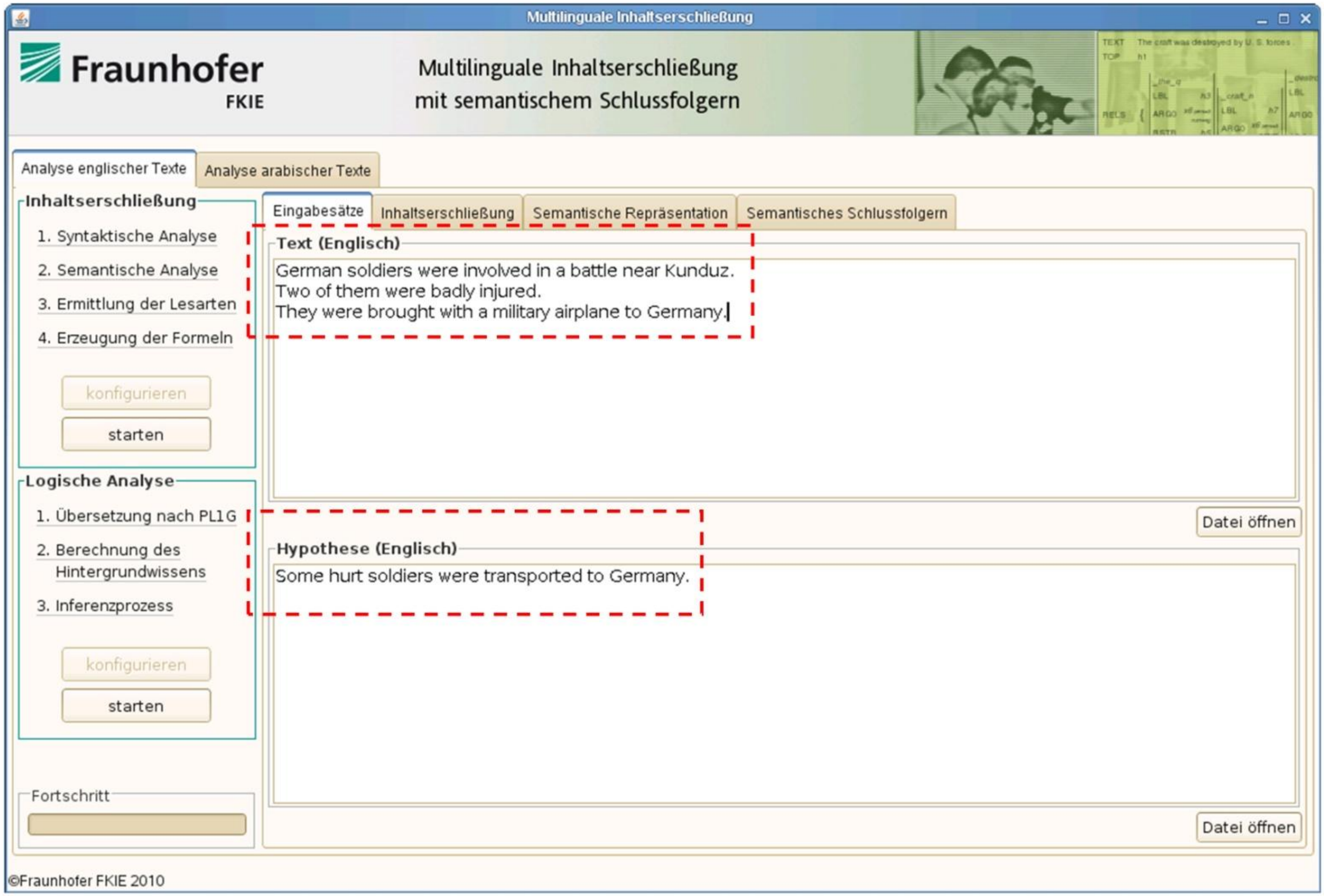
- Text T:
  German soldiers were involved in a battle near Kundus. Two of them were badly injured. They were brought with a military airplane to Germany.
- Hypothesis H:
  Some hurt soldiers were transported to Germany.

1. Introduction - mIE Project – Logical inference II

- Drawing inferences on military relevant texts can be formulated as a problem of recognizing textual entailment (RTE) - a well known academic problem.
- In RTE we want to identify automatically the type of a logical relation between two input texts (T and H).
- The mIE system can be used to find answers to the following, mutually exclusive conjectures with respect to background knowledge:
  1. T entails H,
  2. $T \wedge H$ is inconsistent, i.e., $T \wedge H$ contains some contradiction, or
  3. H is informative with respect to T, i.e., T does not entail H and $T \wedge H$ is consistent.

1. Introduction - mIE Project – Prototype I

- English input.



Text (Englisch)

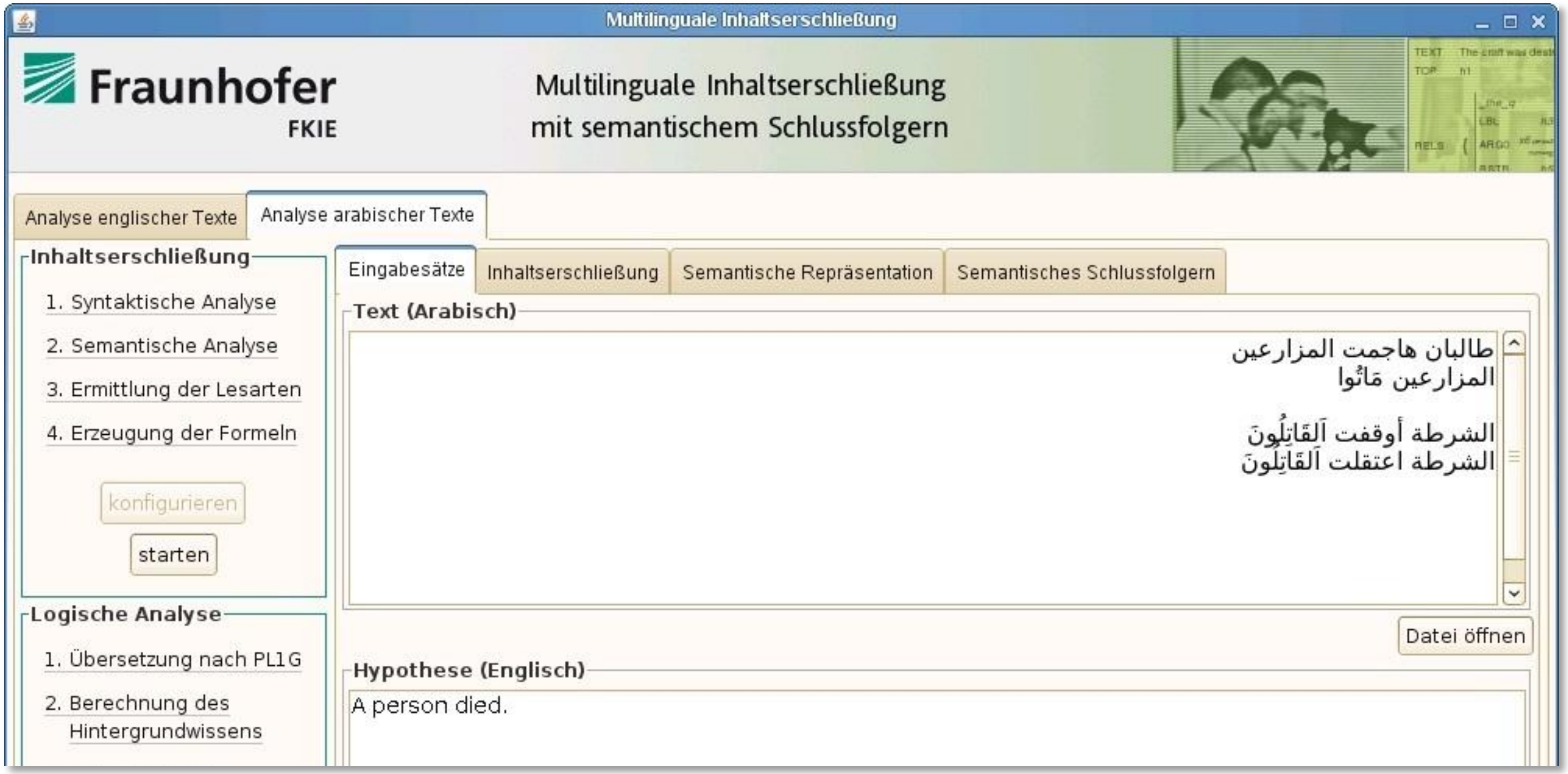
German soldiers were involved in a battle near Kunduz.
Two of them were badly injured.
They were brought with a military airplane to Germany.

Hypothese (Englisch)

Some hurt soldiers were transported to Germany.

1. Introduction - mIE Project – Prototype II

- A second language.



Text (Arabisch)

طالبان هاجمت المزارعين
المزارعين مَاتُوا

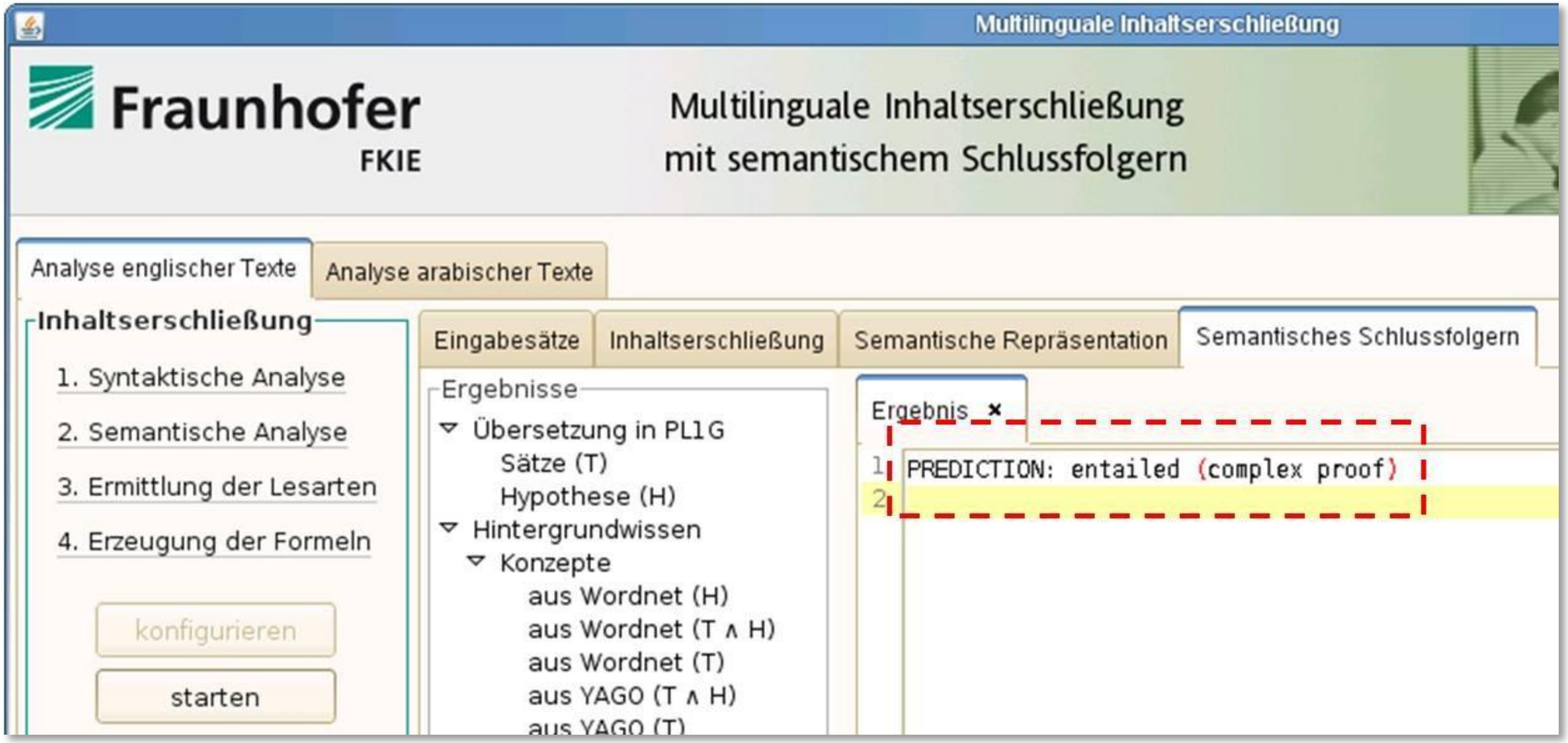
الشرطة أوقفت اَلقَاتِلُونَ
الشرطة اعتقلت اَلقَاتِلونَ

Hypothese (Englisch)

A person died.

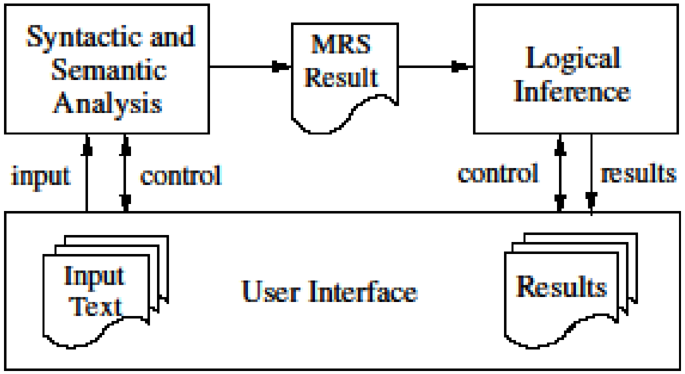
1. Introduction - mIE Project – Prototype III

- Result of the inference process.



```
PREDICTION: entailed (complex proof)
```

1. Introduction - mIE Project – Architecture



Main modules:

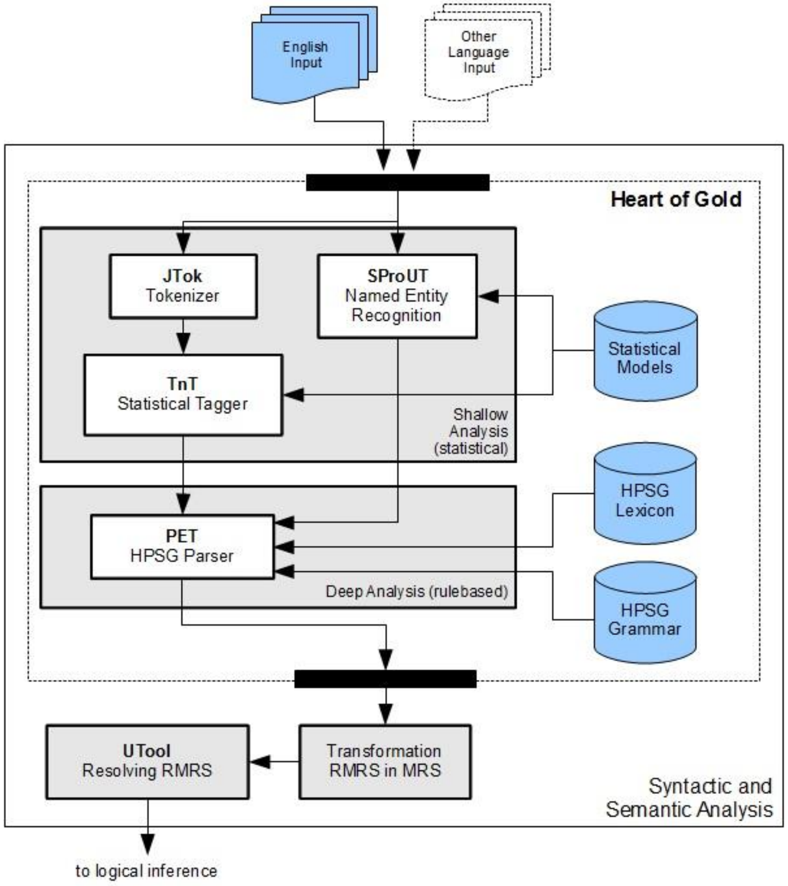
- Syntactic and semantic analysis
- Logical Inference
- Minimal Recursion Semantics (MRS)
- Graphical User Interface (GUI)

2. Combined Deep and Shallow Parsing - I

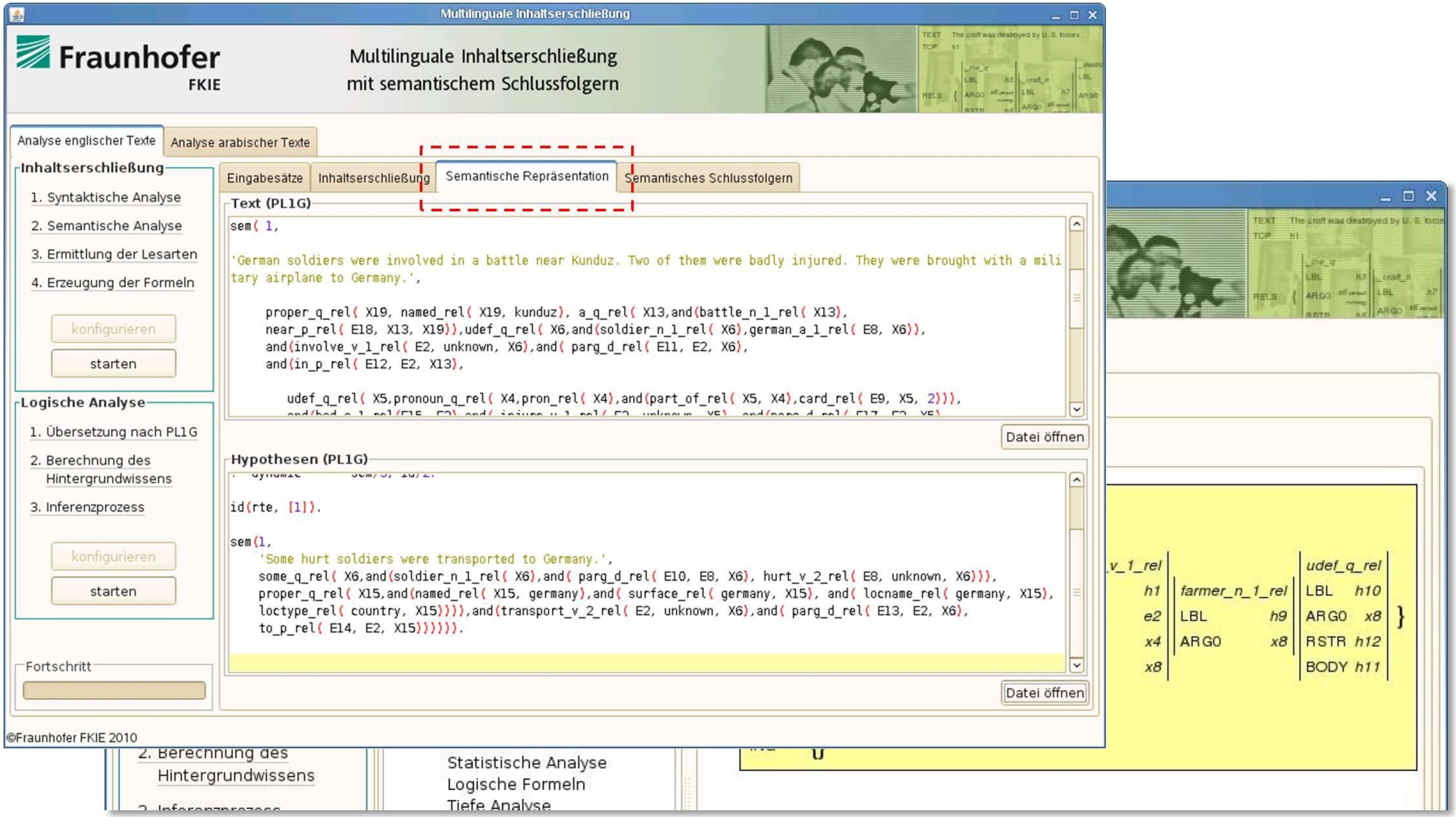
- Task of this module: syntactic processing and semantic construction.
- XML-based middleware architecture Heart of Gold.
- Flexible integration of shallow and deep linguistics-based and semantics-oriented NLP components.
- Shallow processing: statistical or simple rule-based, typically finite-state methods.
- Deep HPSG parser PET.
- English Resource HPSG Grammar (ERG); simple Arabic HPSG grammar.

2. Combined Deep and Shallow Parsing - II

- Tokenization: Java tool Jtok.
- Part-of-speech tagging: statistical tagger TnT trained for English on the Penn Treebank.
- Named entity recognition: SProUT.
- HPSG parser PET: highly efficient runtime parser for unification-based grammars; core of the rule-based, fine-grained deep analysis.
- Robust Minimal Recursion Semantics (RMRS).



2. Combined Deep and Shallow Parsing - III



Text (PL1G)

```
sem( 1,

'German soldiers were involved in a battle near Kunduz. Two of them were badly injured. They were brought with a mili
tary airplane to Germany.',

    proper_q_rel( X19, named_rel( X19, kunduz), a_q_rel( X13,and(battle_n_1_rel( X13),
    near_p_rel( E18, X13, X19)),udef_q_rel( X6,and(soldier_n_1_rel( X6),german_a_1_rel( E8, X6)),
    and(involve_v_1_rel( E2, unknown, X6),and( parg_d_rel( E11, E2, X6),
    and(in_p_rel( E12, E2, X13),

       udef_q_rel( X5,pronoun_q_rel( X4,pron_rel( X4),and(part_of_rel( X5, X4),card_rel( E9, X5, 2))),
```

Hypothesen (PL1G)

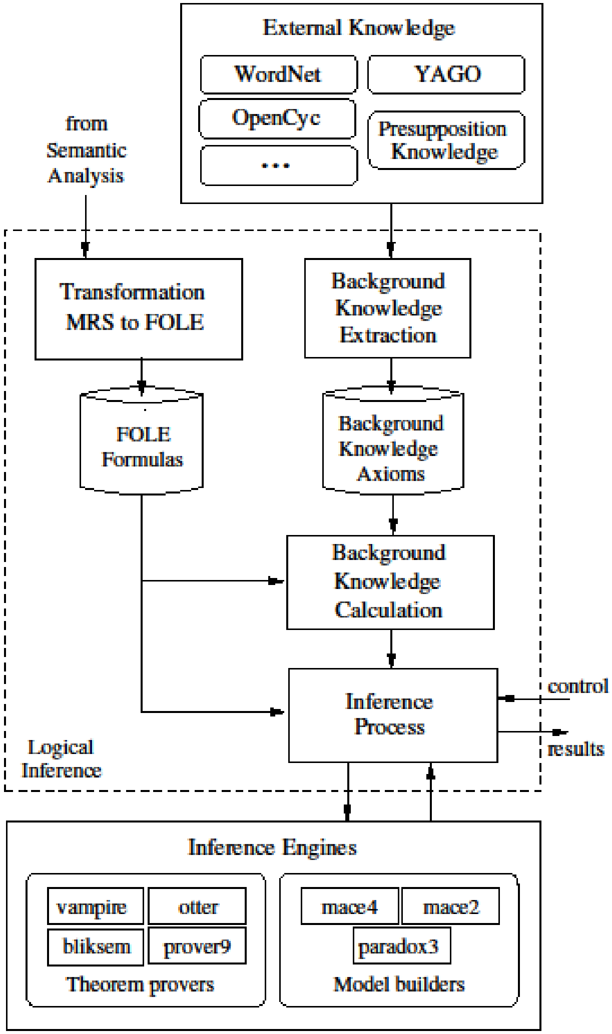
```
id(rte, [1]).

sem(1,
   'Some hurt soldiers were transported to Germany.',
   some_q_rel( X6,and(soldier_n_1_rel( X6),and( parg_d_rel( E10, E8, X6), hurt_v_2_rel( E8, unknown, X6))),
   proper_q_rel( X15,and(named_rel( X15, germany),and( surface_rel( germany, X15), and( locname_rel( germany, X15),
   loctype_rel( country, X15)))),and(transport_v_2_rel( E2, unknown, X6),and( parg_d_rel( E13, E2, X6),
   to_p_rel( E14, E2, X15)))))).
```

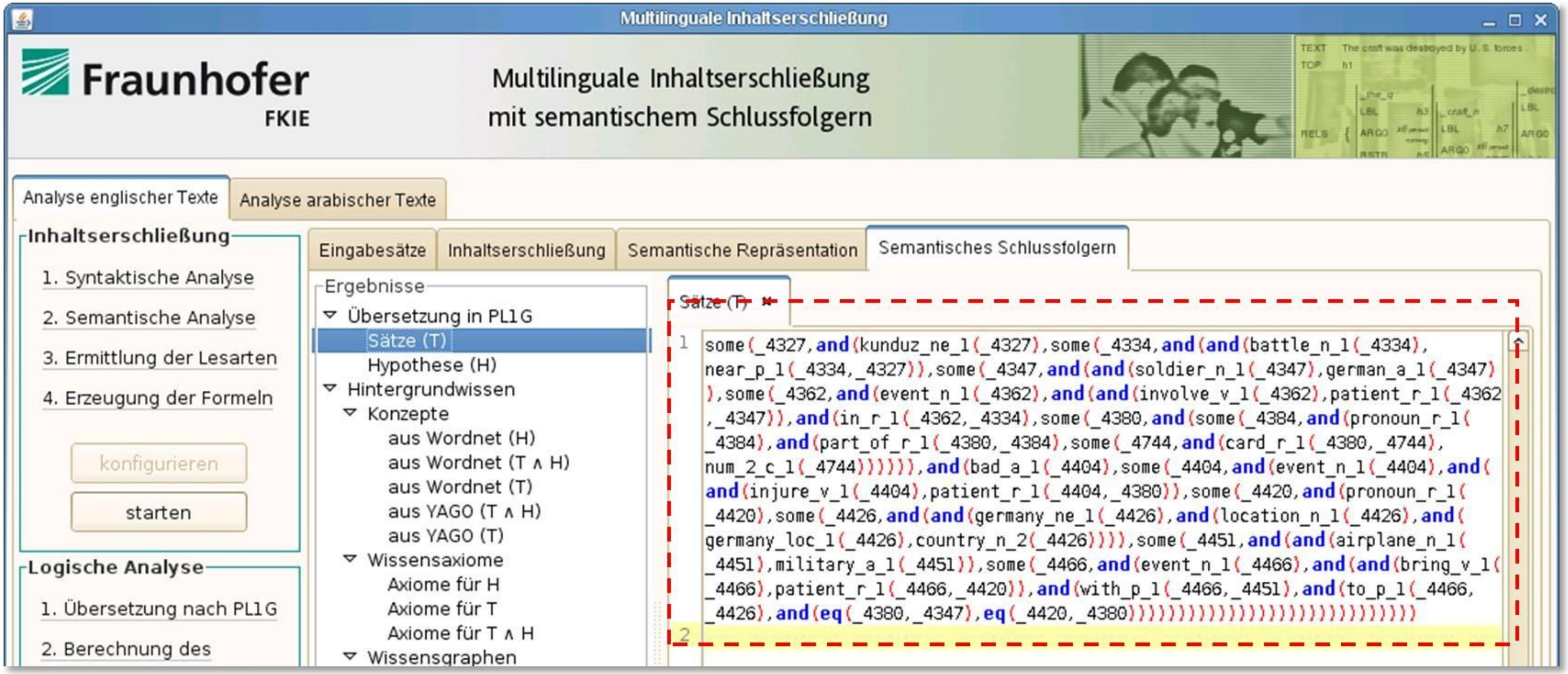
- Result of the combined deep and shallow parsing.

3. Logical Inferences on Text Content - I

- Task of this module: logical deduction, integration of background knowledge.
- The MRS expressions are translated into a semantic equivalent representation of First-Order Logic with Equality (FOLE).
- Find the relevant background knowledge.
- Inference engines:
  - Theorem provers: prove that a formula is valid.
  - Model builders: show that a formula is true in at least one model.
  - The theorem prover attempts to prove the input whereas the model builder simultaneously tries to find a model for the negation of the input.



3. Logical Inferences on Text Content - II



```
some(_4327,and(kunduz_ne_1(_4327),some(_4334,and(and(battle_n_1(_4334),
near_p_1(_4334,_4327)),some(_4347,and(and(soldier_n_1(_4347),german_a_1(_4347)
),some(_4362,and(event_n_1(_4362),and(and(involve_v_1(_4362),patient_r_1(_4362
,_4347)),and(in_r_1(_4362,_4334),some(_4380,and(some(_4384,and(pronoun_r_1(
_4384),and(part_of_r_1(_4380,_4384),some(_4744,and(card_r_1(_4380,_4744),
num_2_c_1(_4744)))))),and(bad_a_1(_4404),some(_4404,and(event_n_1(_4404),and(
and(injure_v_1(_4404),patient_r_1(_4404,_4380)),some(_4420,and(pronoun_r_1(
_4420),some(_4426,and(and(germany_ne_1(_4426),and(location_n_1(_4426),and(
germany_loc_1(_4426),country_n_2(_4426)))),some(_4451,and(and(airplane_n_1(
_4451),military_a_1(_4451)),some(_4466,and(event_n_1(_4466),and(and(bring_v_1(
_4466),patient_r_1(_4466,_4420)),and(with_p_1(_4466,_4451),and(to_p_1(_4466,
_4426),and(eq(_4380,_4347),eq(_4420,_4380)))))))))))))))))))))))))))))
```

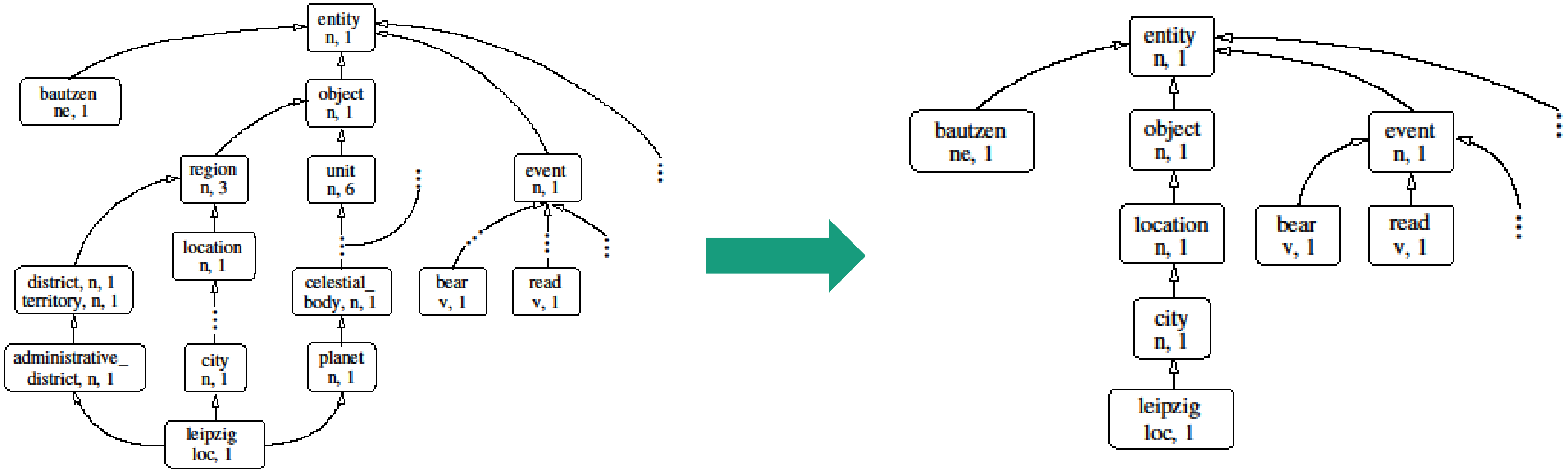
- Semantic representation of T as a FOLE formula.

4. Background Knowledge - I

- Extend automatically the FOLE formulas (T and H) with problem-relevant knowledge in form of background knowledge axioms.
- 1st source: WordNet 3.0
  - A lexical database for synonymy, hyperonymy (e.g., *location* is a hyperonym of *city*), and hyponymy (e.g., *city* is a hyponymy of *location*) relations (taxonomy).
  - Approx. 2.6 million entries.
  - It helps the logical inference process to detect entailments between lexical units from the text and the hypothesis.
  - The hyperonymy/hyponymy relation in WordNet spans a directed acyclic graph (DAG) with the root node ‘entity’ => may induce inconsistencies between the input problem formulas and the extracted knowledge. This must be taken into account during the integration process.

4. Background Knowledge - II

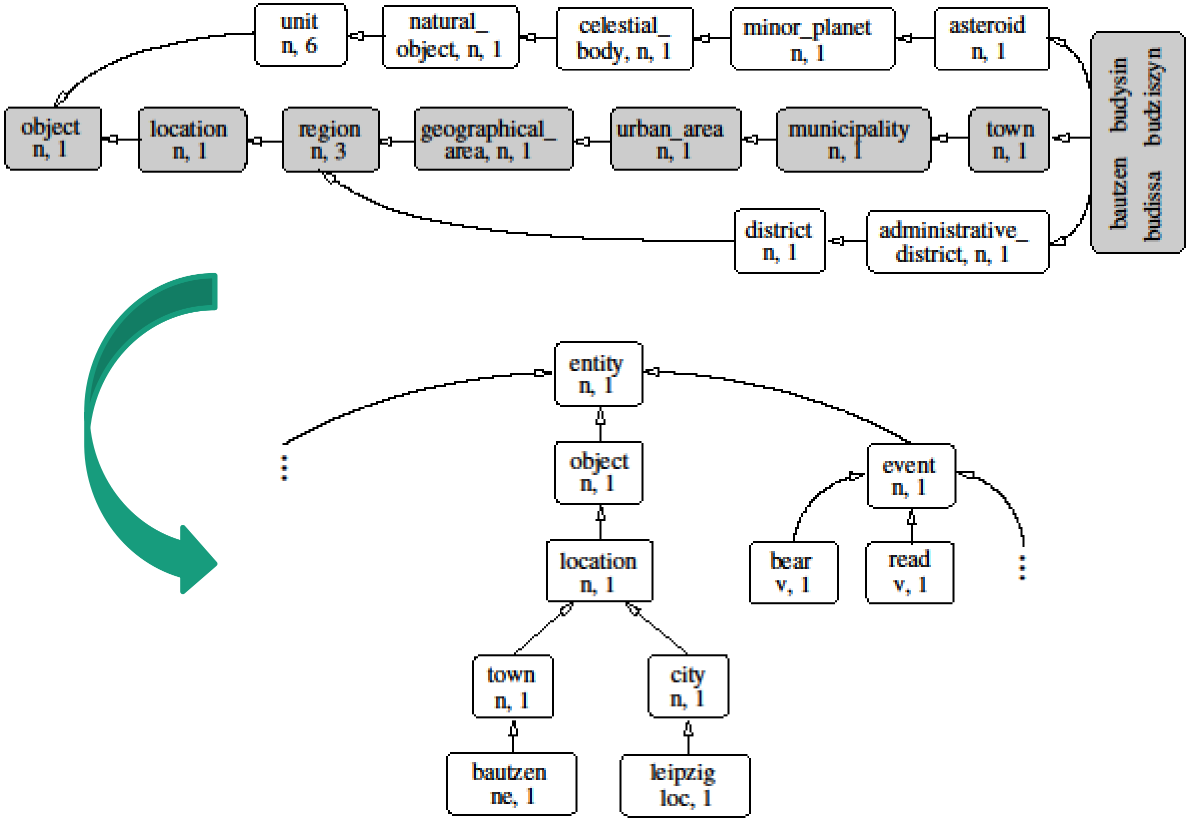
- Integration of WordNet
  - List all concepts and individuals from the input formulas.
  - Find the search predicates in WordNet and build the knowledge graph (using hyperonymy/hyponymy and synonymy relations).
  - The graph is optimized so that only those concepts appear in a tree, which are directly relevant for the inference problem.



4. Background Knowledge - III

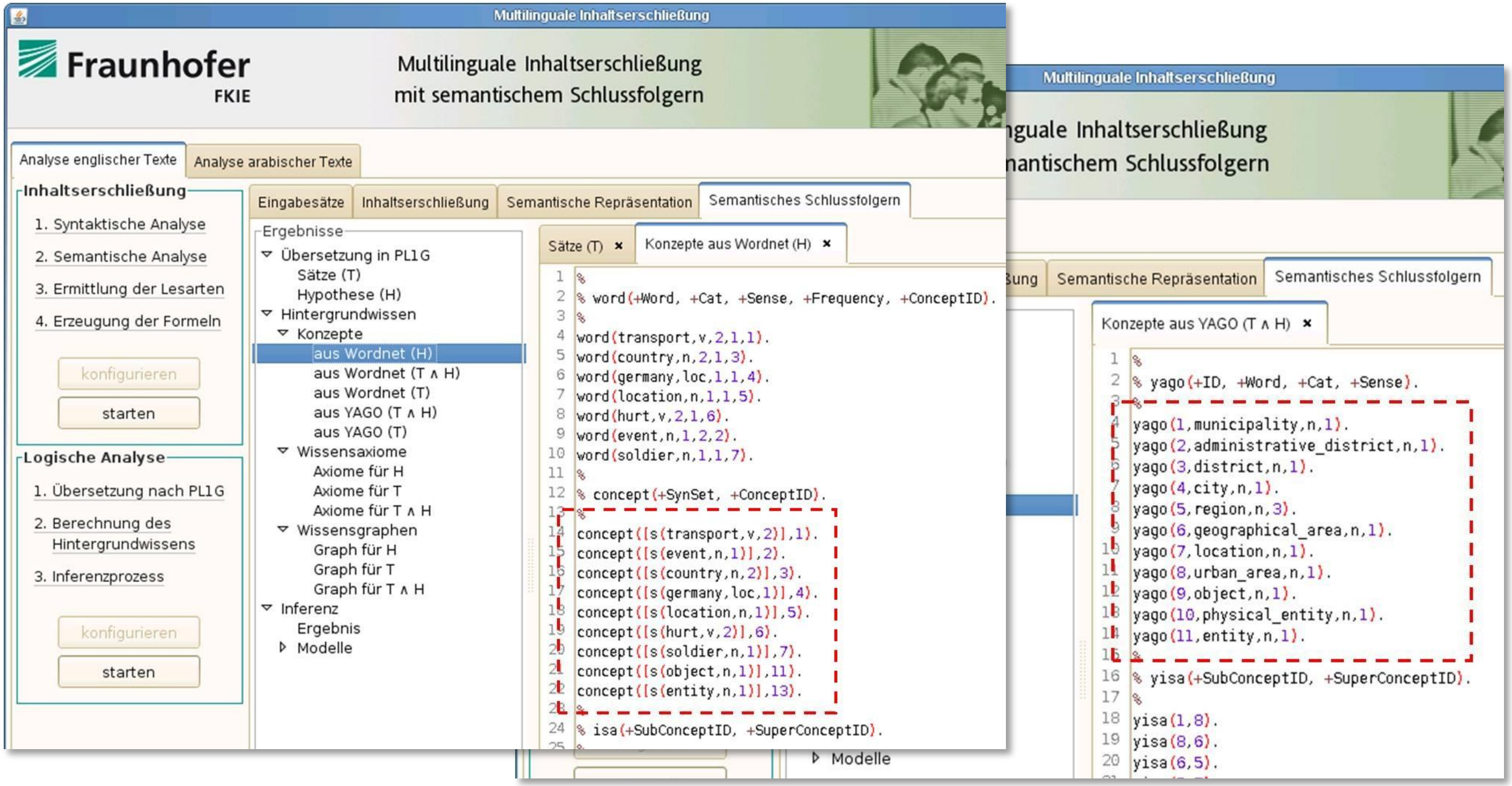
- 2nd source: YAGO
  - Large ontology; approx. 22 million facts and relations.
  - Assembled automatically from the category system and the info boxes of Wikipedia, and combined with taxonomic relations from WordNet.
- Integration of YAGO
  - Consult YAGO about search predicates that were not recognized in the WordNet phase.
  - The result of every YAGO-query is in general represented by a DAG.
  - Preserve correctness of results: select for the integration only those concepts, individuals, and relations which are on the longest path from the most general concept to one of the direct hyperonyms of the leaf.

4. Background Knowledge - IV



- Result of a query to YAGO and integration of the result.

4. Background Knowledge - V



```
%
% word(+Word, +Cat, +Sense, +Frequency, +ConceptID).
%
word(transport,v,2,1,1).
word(country,n,2,1,3).
word(germany,loc,1,1,4).
word(location,n,1,1,5).
word(hurt,v,2,1,6).
word(event,n,1,2,2).
word(soldier,n,1,1,7).
%
% concept(+SynSet, +ConceptID).
%
concept([s(transport,v,2)],1).
concept([s(event,n,1)],2).
concept([s(country,n,2)],3).
concept([s(germany,loc,1)],4).
concept([s(location,n,1)],5).
concept([s(hurt,v,2)],6).
concept([s(soldier,n,1)],7).
concept([s(object,n,1)],11).
concept([s(entity,n,1)],13).
%
% isa(+SubConceptID, +SuperConceptID).
%
```

```
%
% yago(+ID, +Word, +Cat, +Sense).
%
yago(1,municipality,n,1).
yago(2,administrative_district,n,1).
yago(3,district,n,1).
yago(4,city,n,1).
yago(5,region,n,3).
yago(6,geographical_area,n,1).
yago(7,location,n,1).
yago(8,urban_area,n,1).
yago(9,object,n,1).
yago(10,physical_entity,n,1).
yago(11,entity,n,1).
%
% yisa(+SubConceptID, +SuperConceptID).
%
yisa(1,8).
yisa(8,6).
yisa(6,5).
```

- Concepts from WordNet and YAGO.

5. Conclusion

- In this presentation, we introduced the mIE system based on a combination of deep and shallow parsing with logical inferences on the analysis results and background knowledge.
- Possible improvements
  - The Arabic HPSG grammar is only a very small one.
  - During the inference process only the most probable meaning of the words is considered. Considering as well other - less probable - meanings might increase the inferential power.
  - It would be interesting to look at the inconsistent cases of the inference process. They were caused by errors in presupposition and anaphora resolution, incorrect syntactic derivations, and inadequate semantic representations.
  - For the implementation of some temporal calculus, also temporal relations from YAGO such as *during*, *since*, or *until* could be considered.
  - …

5. References

- A. Wotzlaw and R. Coote. *Recognizing textual entailment with deep-shallow semantic analysis and logical inference*. In: SEMAPRO 2010, Florence, Italy, 2010.
- R. Coote and A. Wotzlaw. *Generation of first-order expressions from a broad coverage HPSG grammar*. In AAIA'10, Wisla, Poland, 2010.
- Andreas Wotzlaw. *Towards better ontological support for recognizing textual entailment*. In: EKAW 2010, Lisbon, Portugal, 2010.
- M. Hecking, A. Wotzlaw, R. Coote. *Abschlussbericht des Projektes Multilinguale Inhaltserschließung*. FKIE-Bericht Nr. 207, Wachtberg, Germany, 2011.
- M. Hecking. *Multilinguale Textinhaltserschließung auf militärischen Texten*. In: Verteilte Führungsinformationssysteme. M. Wunder, J. Grosche (Hrsg.), Springer-Verlag, 2009.
- M. Hecking and T. Sarmina Baneviciene. *A Tajik Extension of the Multilingual Information Extraction System ZENON*. In Proceedings of the 15th International Command and Control Research and Technology Symposium (ICCRTS), Santa Monica, CA, U.S.A., 2010.
- M. Hecking. *System ZENON – Semantic Analysis of Intelligence Reports*. In: Proceedings of the LangTech 2008, February 28-29, 2008, Rome, Italy.

Thank you for your attention!

Questions?