Multilingual Information Extraction for Intelligence Purposes

Christina Schwerdt
Dr. Matthias Hecking
Forschungsgesellschaft für Angewandte Naturwissenschaften e.V. (FGAN)
Forschungsinstitut für Kommunikation, Informationsverarbeitung und Ergonomie (FKIE)
Abtl. Informationstechnik und Führungssysteme (ITF)
Neuenahrer Straße 20
53343 Wachtberg-Werthhoven
hecking@fgan.de

Content

- 1. Introduction
- 2. The ZENON System Version 1
- 3. Multilingual Information Extraction
  - Multilingual IE in the ZENON system
  - Dari processing resources and named entity recognition
  - Dari-German dictionary
  - Morphological analysis of verb groups
  - Open problems
- 4. Conclusion

1 | Introduction - I

- Processing of human language as a critical capability in many future military applications (cf. [Steeneken, 1996]).
- Content analysis/extraction of free-form texts is important for any information operation of the Network Centric Warfare (NCW) concept (cf. [NCW, 2001], p. 5-15).
- This can be realized through Information Extraction (IE) which is a natural language processing technique (cf. [Appelt, 1999], [Hecking, 2004a]).

1 | Introduction - II

- Specific problem: content extraction of intelligence reports
- ZENON project version 1: The overall objective is to realize experimental systems for (partial) *content extraction of HUMINT reports* from the KFOR deployment of the Bundeswehr and to realize a possibility to evaluate the formal representation of the content.
- For the realization of the IE module 4,498 English HUMINT reports are available; KFOR corpus: 800 annotated reports
- ZENON system version 1 can only process documents in English

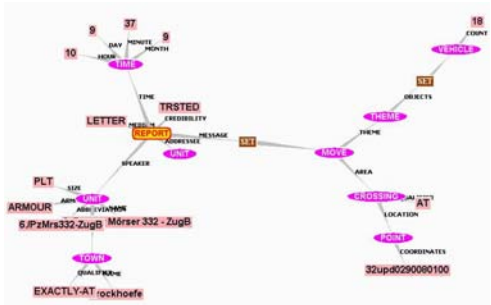
1 | Introduction - III

- ZENON system version 2: IE result from two different languages will be combined; ongoing work
- This will allow the analyst to access information from texts written in different languages without knowing these languages.
- In version 2 the language *Dari* is used as the second one.
- Two official languages of Afghanistan: Dari, Pashtu; operational area of the German Bundeswehr
- Also: a word-to-word-translation to further support the analyst.

Dr. M. Hecking

The ZENON version 1 research system:

- The Information Extraction (IE) technology is used for the content extraction.
- The information about the actions and named entities are identified from each sentence and the content of the sentences are formally represented in typed feature structures.
- These structures can be combined and presented in a graphically navigatable *Entity-Action-Network*.



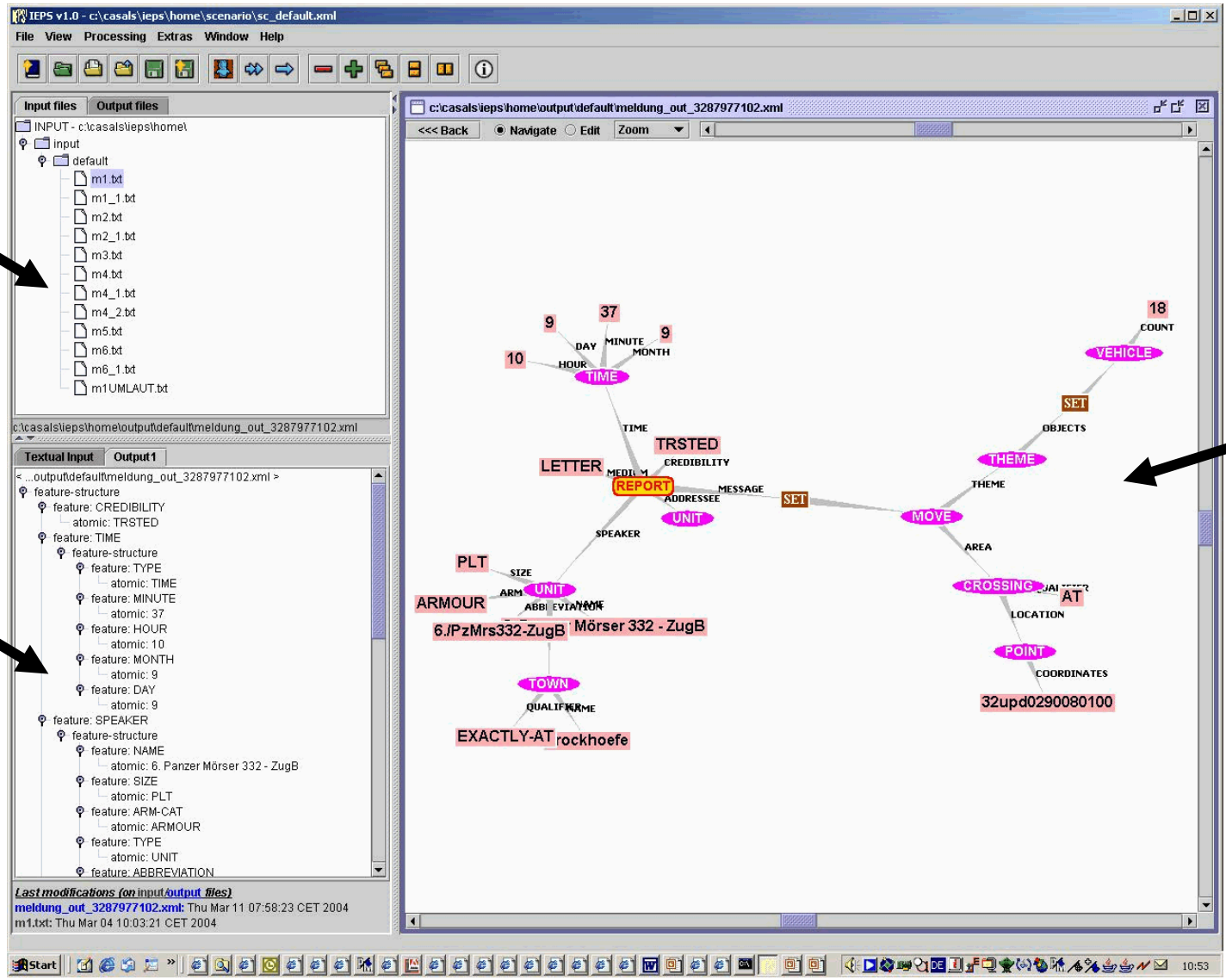
2 | ZENON System Vers. 1 - III

Dr. M. Hecking

input files

formal output

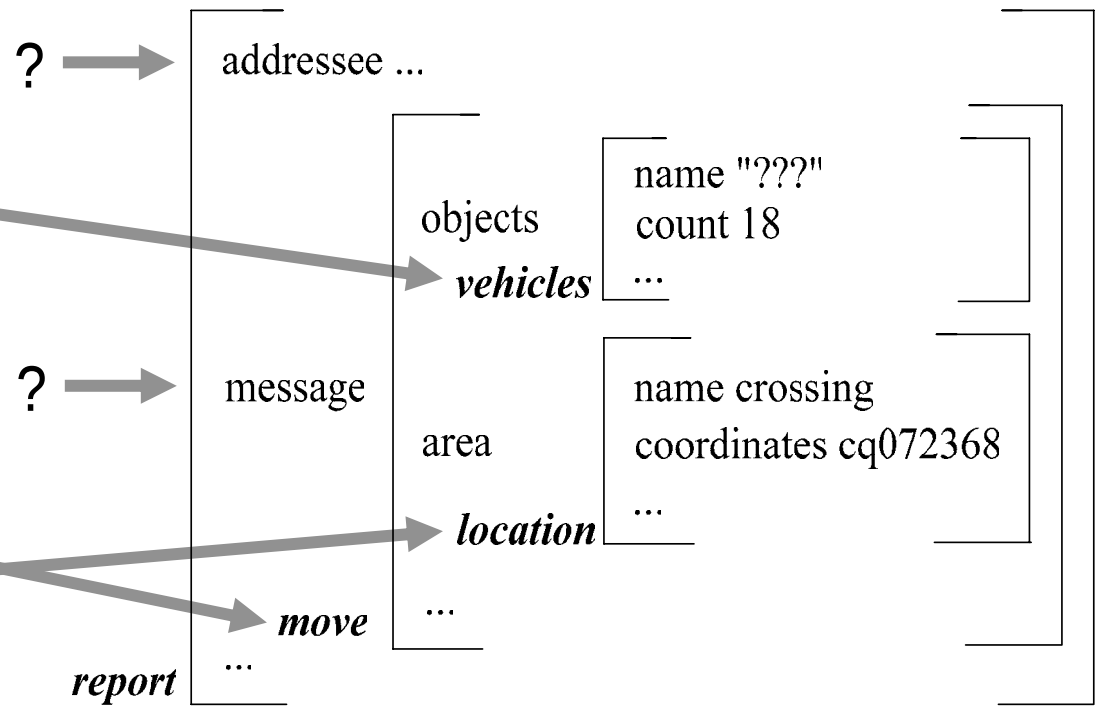
graphical output



Dr. M. Hecking

„... in TUZLA:
10.20 a.m.
18 vehicles
- 8 with attached ZIS–3 and 1 with attached T 12 - march

at road crossing (CQ 072368) south of MILESKIJ (CQ 0737) to the north.“

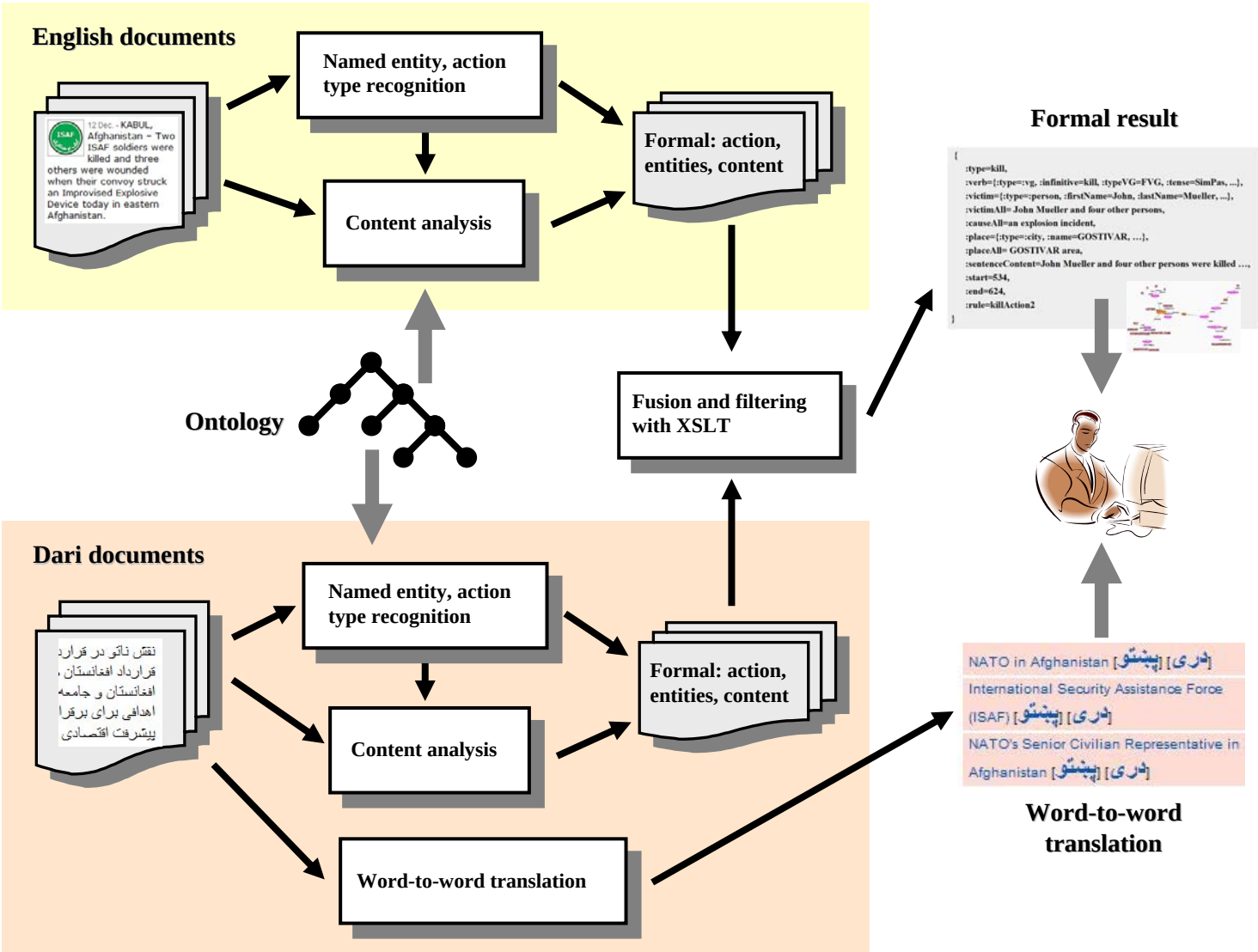


- Information extraction (IE) is the task of identifying, collecting and normalizing information from natural language text.
- Relevant information about the Who, What, When, etc. is looked for.
- The information of interest is described through domain-specific lexicon rules and patterns called *templates*.
- During the IE task these templates are filled with the collected information.
- The templates are domain and task specific, i.e. for each new task and domain they must be newly created.

- Technical basis of the information extraction:
  - extensive linguistic knowledge
    - general and application-specific lexica
    - general and/or application-specific phrase grammars
    - general (and/or application-specific) clauses and sentence grammars
  - cascaded transducers, i.e. finite state automaton that reads from the input and writes to the output
  - only application-relevant parts of the texts are analyzed through transducers (shallow parsing techniques)

- Multilingual (as well as multi-source) information extraction is an actual research topic
- Main idea of multilingual information extraction: extraction of information about a specific entity and/or action from documents written in different languages
- Hypothesis: using information written in different languages gives more accurate intelligence results

3 | Multilingual IE - Main idea

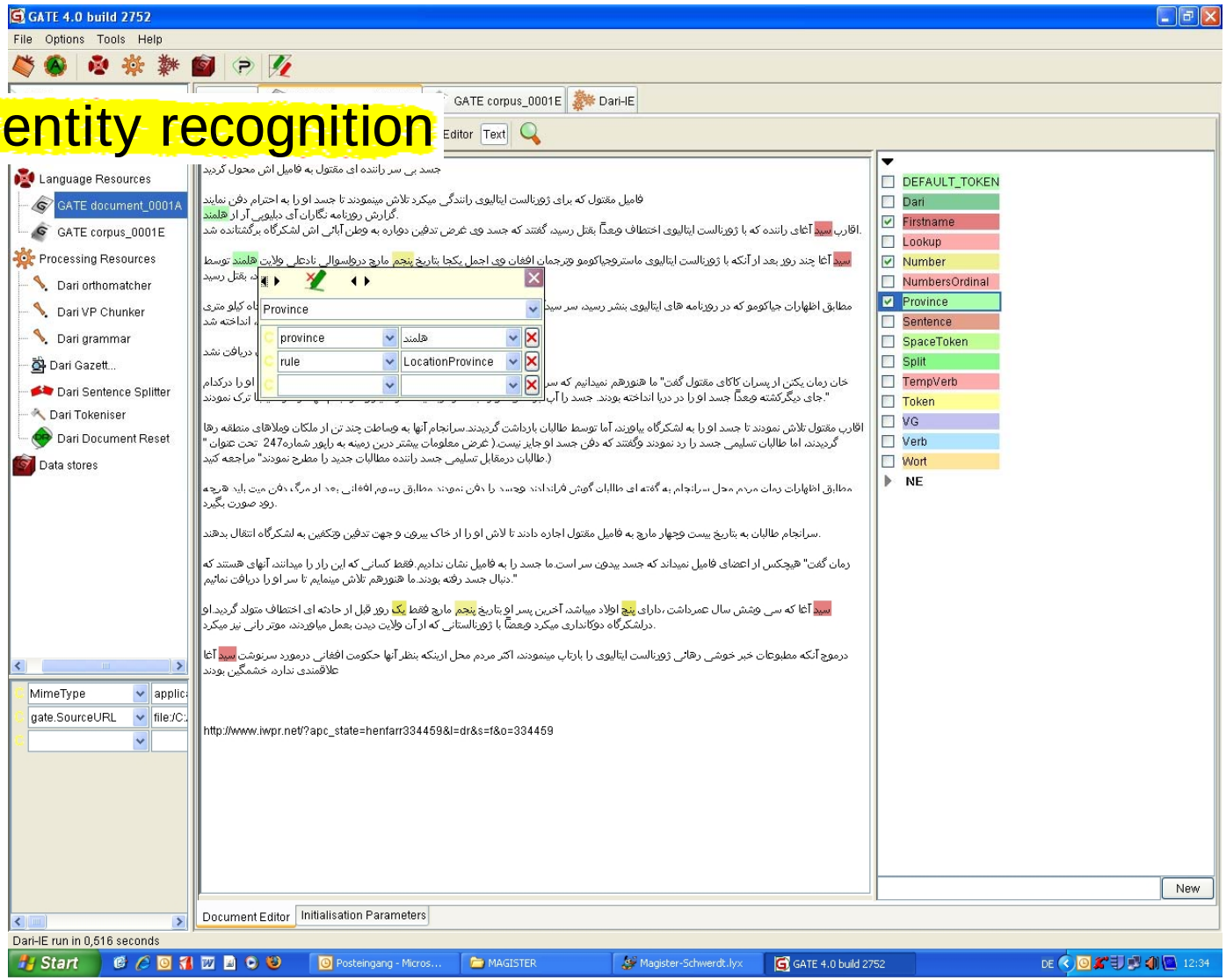


- For the Dari language natural language processing functionality was realized in a master thesis (cf. [Schwerdt, 2007a]).
- Components of the Dari application: *Dari Document Reset, Dari Tokenizer, Dari Sentence Splitter, Dari Gazetteer, Dari Grammar, Dari VP Chunker*.
- Dari Gazetteer (lists of names): *city_afgh* (28), *city_world* (126), *country_world* (171), *days* (7), *female_names* (708), *male_names* (829), *months* (12), *numbers* (11), *ordinals* (10), *province_afgh* (34).
- Rules for named entity recognition

3 | Multilingual IE - Dari processing ... - II

Dr. M. Hecking

Named entity recognition



Dr. M. Hecking

- A Dari-German full-form dictionary with approximately 48,000 entries was integrated in the Gazetteer
- For each dictionary entry the following information is available: *ID, source of the entry, subject area, German translation, German part of speech, German grammatical gender, Number of words in German, Dari lemma/headword, Dari number, English translation, source of the Dari word, Dari part of speech, Number of words in Dari, Dari dialect, description, frequency*

Dictionary entries

ball

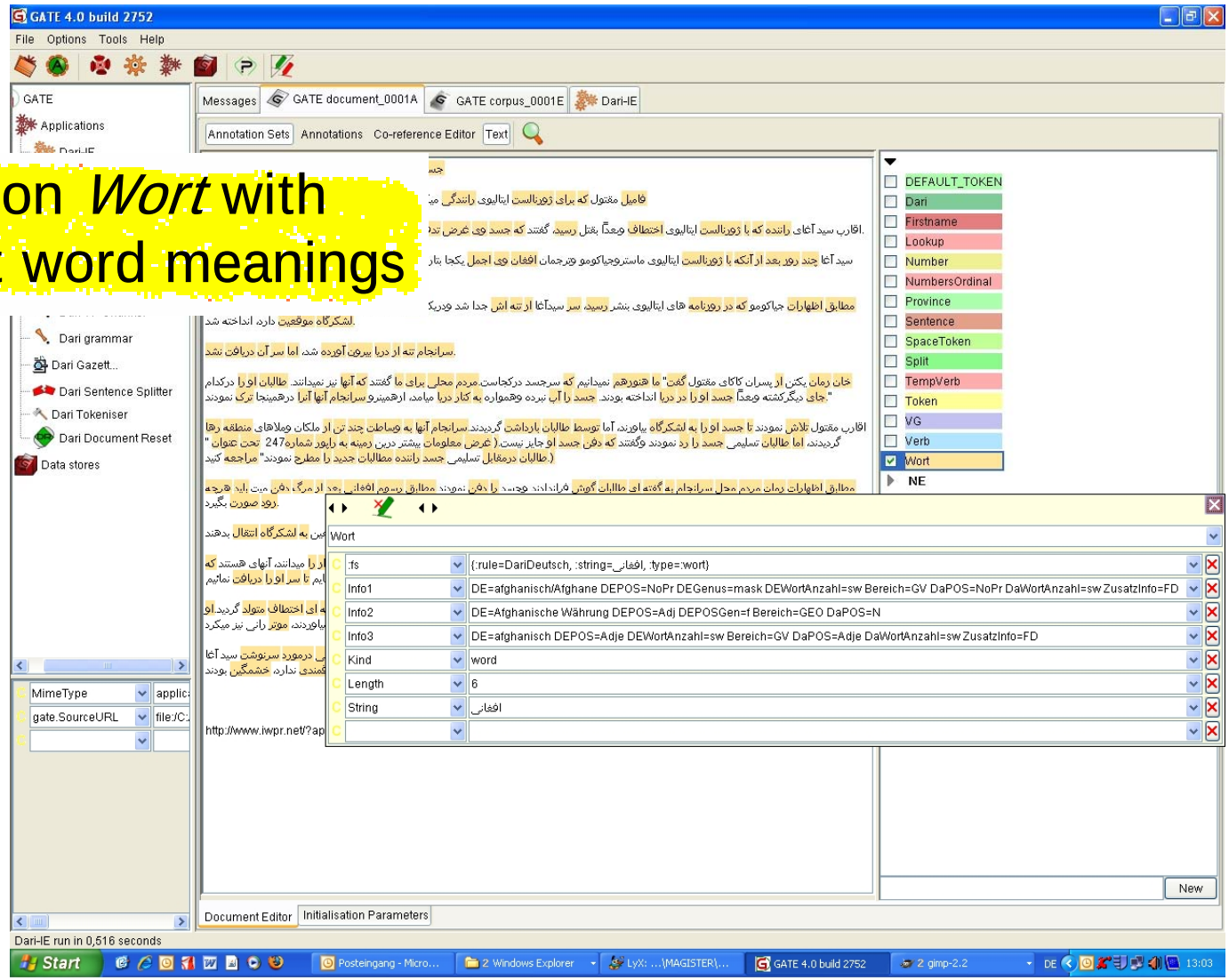
ID	60532	60544
Herkunft	Dr_De_xxx200611	DR_DE_xxx2006-12-34974
Sachgebiet	GV	GV
Dt. Übersetzung	Ball	Schnecke
Dt. Wortart	N	NoCo
Dt. Genus (+Wortart)	N,m	
Anzahl d. dt. Wörter		sw
Dt. Genus		fem
Dari-Lemma	گوی	گوک
Quelle		LTWB
Dari Wortart	N	NoCo
Dari Wortanzahl		sw
Dialekt		FD

snail

3 | Multilingual IE - Dari-German dictionary - III

Dr. M. Hecking

Annotation *Wort* with different word meanings



Dr. M. Hecking

- For the identification of the action types used in the Dari documents the verbal phrases must be analyzed.
- Morphological analysis for the most important verbs in Dari
- Two full-form lexica
  - 39,552 inflectional past tense forms of the 8,818 verbs from the Dari-German dictionary
  - 6,592 past-participle forms of the verbs

Dr. M. Hecking

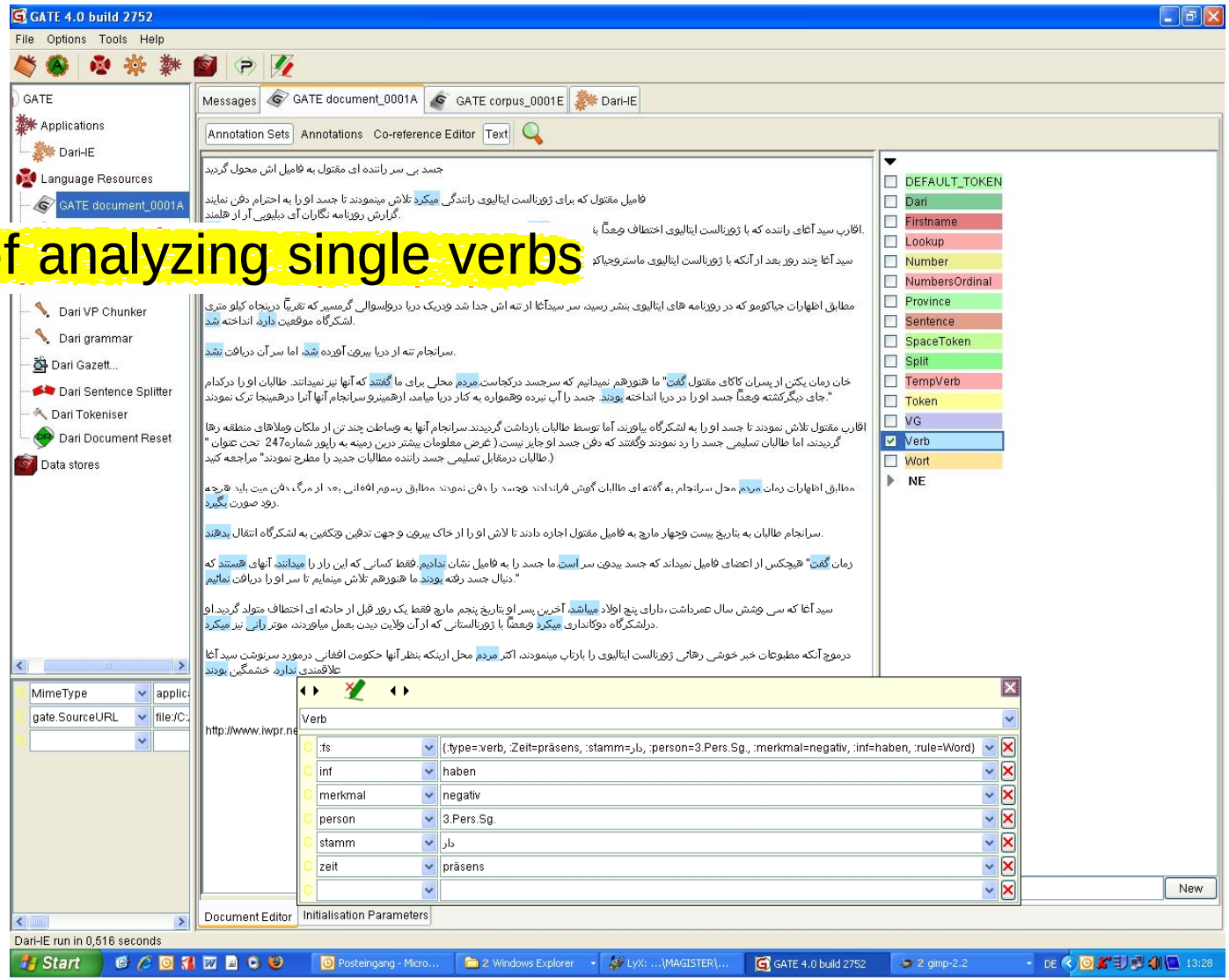
- The forms of the present tense in Dari are irregular.
- The morphological analysis of the present tense is restricted (at the moment) to the 84 most often used verbs.
- Information provided by the morphological analysis: *stem*, *mood* (indicative, subjunctive, and imperative), *number*, *tense*, *negation*, *voice* and *translation of the infinitive*
- Compound tenses are also handled.

Past tense forms of *budan* ("to be")

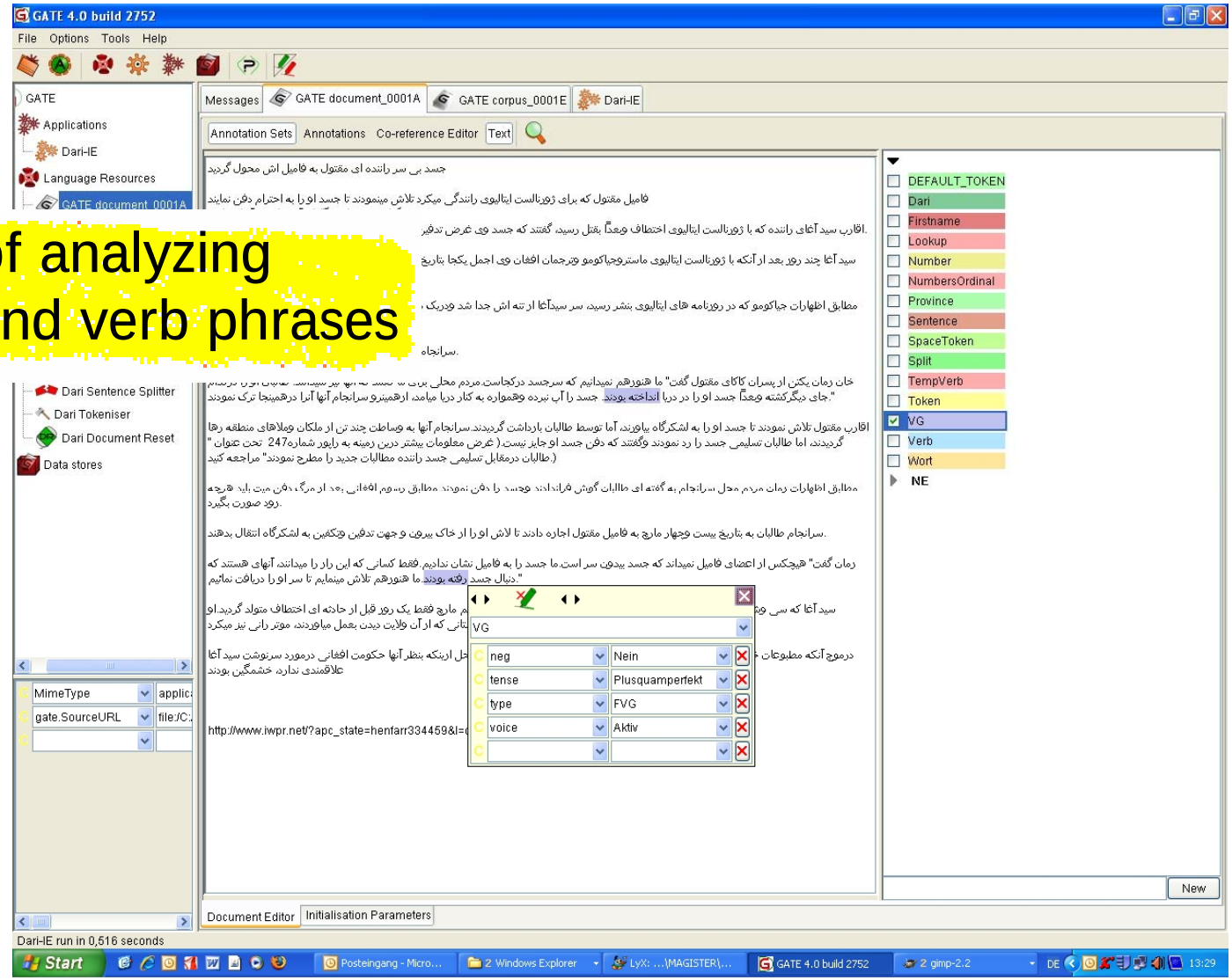
			->	budam بودم
			->	bude بودی
budan بودن	->	bud بود	->	bud بود
			->	budem بودیم
			->	buded بودید
			->	budand بودند

Dr. M. Hecking

Result of analyzing single verbs



Dr. M. Hecking



Result of analyzing compound verb phrases

- Transliteration of Dari names to those written in roman scripts.
- Spelling variants in both languages; at the moment only an ad hoc solution; more principal solution needed: for each name *one* canonical form and its spelling variants are defined
- Tight integration of the word-to-word translation with the information extraction result
- How to get the same coverage of linguistic data (dictionaries, grammars ...) and functionality (POS tagger, morphology ...) for less commonly spoken languages?

4 | Conclusion

- ZENON system version 1 explained
- Main idea of multilingual information extraction
- 2nd language: Dari
- Named entity recognition for Dari texts
- Integration of Dari-German dictionary
- Morphological analysis of Dari verb groups
- Open problems
- We are in the process of integrating the modules for the semantic analysis of English and Dari documents.