13th ICCRTS
"C2 for Complex Endeavors"

June 17-19, 2008
The Meydenbauer Center
Bellevue, WA
U.S.A.

Paper's Title: Multilingual Information Extraction for Intelligence Purposes

Topic: Topic 1: C2 Concepts, Theory, and Policy
Topic 11: Multinational Endeavors

Authors

Name:	Dr. Matthias Hecking (point of contact)
	Christina Schwerdt
Organization:	FGAN/FKIE
Address:	Neuenahrer Straße 20
	53343 Wachtberg-Werthhoven
	Germany
Phone:	+49 228 9435 576
Fax:	+49 228 9435 685
E-Mail:	hecking@fgan.de

Multilingual Information Extraction for Intelligence Purposes

Dr. Matthias Hecking, Christina Schwerdt
FGAN/FKIE
Neuenahrer Straße 20
53343 Wachtberg-Werthhoven, Germany
hecking@fgan.de

Abstract

The new deployments of the German Federal Armed Forces cause the necessity to analyze large quantities of intelligence reports and other documents written in different languages. For the monolingual information extraction we have build the ZENON system, which can be used for the (partial) content analysis of English HUMINT reports from the KFOR deployment of the Bundeswehr. In this paper, we present how we extended the ZENON system to do information extraction from texts written in *different* languages. The general hypothesis behind this work is the claim that information extraction on texts written in different languages yields better intelligence results. We present the main idea of multilingual information extraction and show, how named entity recognition was realized for Dari texts. A Dari-German dictionary extends the capability of the information extraction functionality. This is presented, as well as the morphological analysis of Dari verb groups. After a short introduction, the information extraction and the ZENON project are explained. In the main part of the paper, our approach to realize multilingual information extraction is described in detail. First, we explain our main idea. Then, we show named entity recognition for Dari texts, the integration of the Dari-German dictionary and the morphological analysis of the verb groups. Finally, open problems are discussed.

1. Introduction

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

We set up the *research project ZENON*[1], in which the IE approach is used for the (partial) content analysis of English HUMINT reports from the KFOR deployment of the Bundeswehr (cf. [Hecking, 2007a], [Hecking, 2006a], [Hecking, 2006b], [Hecking, 2005a], [Hecking, 2004a], [Hecking, 2004b], [Hecking, 2003a], [Hecking, 2003b]). The overall objective of this research is to create a *graphically navigatable Entity-Action-Network*. The information about the actions and named entities are identified from each sentence and the content of the sentences are formally represented. These formal representations can be combined and presented in the navigatable network.

The ZENON system in version 1 can only process documents in one language (English). We are in the process of building version 2. In this version of the system the information extraction result from two different languages will be combined. This will allow the analyst to access information from texts written in different languages without knowing these languages.

[1] according to: Zenon of Citium, 336 BC - 264 BC, philosopher, founder of the Stoicism

The general hypothesis behind this work is the claim that information extraction on texts written in different languages yields better intelligence results. In version 2 the language *Dari* is used as the second one. Beside the information extraction the system will give a word-to-word-translation to further support the analyst.

There are two official languages of Afghanistan according to the Constitution. These two official languages are also the most commonly spoken. Dari - a form of the Persian language - is spoken by half of the population. Pashtu is the other official language and is spoken by 35% of the population. We choose Dari for our multilingual approach because this language is spoken in the operational area of the German Bundeswehr.

This paper is structured as follows. First, a short introduction into the ZENON system version 1 is given. In the main part of the paper, the multilingual information extraction approach in the ZENON system version 2 is described. For this, the general concept and the realized Dari functionality are presented. Finally, we discuss open problems.

At the time of writing we are in the process of putting together the already programmed modules.

2. The ZENON System Version 1

Starting with English HUMINT reports (and a list of the city names) from the KFOR deployment of the German Federal Armed Forces we have realized in our research project ZENON a system that is able to do a (partial) content analysis of these reports. The ZENON system uses a *shallow syntactic approach* based on chunk-parsing and transducer. The transducers are handcrafted grammars processed as finite automaton (cf. [Hecking, 2006a], [Hecking, 2006b], [Hecking, 2005a], [Hecking, 2003a], [Hecking, 2003b], [Hecking, 2004a], [Hecking, 2004b]). The content of these KFOR reports are from a wide spectrum. Apart from descriptions of conflicts between ethnic groups, tensions between political parties, information about infrastructure problems, etc. there are also reports, which concern individuals or other entities. Statements of the form *A meets B*, *A marries C*, *A shoots B*, etc. contains information about activities/events and involved entities. This information, completed with location and time data, is combined into a graphically navigatable Entity-Action-Network (e.g.; with a person in the center of the network). The intelligence analysts can use this network to navigate through the meaning space of the reports.

A widely used toolbox for constructing information extraction systems is GATE (General Architecture for Text Engineering, cf. [Cunningham, 2002]). GATE is an architecture, a free open source framework (SDK) and graphical development environment for Natural Language Engineering and offers a lot of tools, which are used to realize the natural language processing parts of the ZENON prototype (e.g., morphological analyzer, part-of-speech (POS) tagger, pre-defined transducer to recognize English verbal phrases, chunk-parsing). Tools for handling languages which use Arabic characters are also enclosed. In addition to with GATE implemented modules ZENON also has a subsystem for visualizing information extracted from free-form texts.

During the processing of the HUMINT reports the natural language text is tokenized (i.e., find words, numbers, etc.), the sentence boundaries are detected, the part-of-speech (i.e., whether it's a noun, a verb, etc.) is determined, simple names of cities, regions, military organizations etc. are annotated (through the Gazetteer), named entities (i.e., complex names of e.g. political organizations, person names, etc.) are recognized and a morphological analysis of the verbs is

done. The results of this are the annotated sentences of the reports. Then these annotations are used to extract the action type (e.g., 'kill') starting with the verb of the sentence. If the action type is determined the other parts of the sentence (e.g., subject, object, time expressions) are located and formally represented in *typed feature structures*. These structures are coded in XML (Extensible Markup Language) format and represent the output of the natural language part of the ZENON system. In the last step the extracted content of different reports are combined and selected according to predefined XSLT (Extensible Stylesheet Language Transformation) sheets. Then, the result of the analysis is depicted graphically and can be navigated interactively.

An important processing step during the natural language processing is the recognition of the domain- and application-specific named entities. In the ZENON system transducers for the recognition of the following named entities were developed: *City, Company, Coordinates, Country, CountryAdj, Currency, Date, GeneralOrg, MilitaryOrg, Number, Percent, Person, PoliticalOrg, Province, Region, River, Time* and *Title*.

Another important step is the extraction of verb phrases, action types and the sentence content. The ZENON prototype uses various transducers (grammars) to recognize finite and non-finite verbal phrases, modal verb phrases, participles and special composed verb expressions.

For a more detailed description of the above described processing steps see [Hecking, 2006a].

4,498 military reports (mostly in English) from the KFOR deployment of the German Federal Armed Forces were used for the realization of the ZENON prototype. From these reports 800 were manually annotated and form the *KFOR Corpus*[2]. The corpus covers 886,000 tokens and contains the annotations in different layers (cf. [Hecking, 2007a], [Hecking, 2006c]).

3. Multilingual Information Extraction

Multilingual (as well as multi-source) information extraction is an actual research topic (cf. [Poibeau, 2007a]). For example, in [Pouliquen, 2007a] a multilingual multi-document continuously-updated social network is described. From 35,000 news articles (per day) in 32 different languages person names are extracted, name variants are matched and a social network graph is produced that shows which names are mentioned together with other names.

The information about a specific entity or a specific action type extracted by the ZENON system is only from documents written in one language (The KFOR reports are in English). The main idea of *multilingual information extraction* is the extraction of information about a specific entity and/or action from documents written in different languages. The hypothesis is that using information written in different languages gives *more accurate intelligence results*. For the realization of multilingual IE, processing resources are needed for each language. Beside language-specific lexica, grammars, morphological analysis etc. a shared set of concepts is necessary. For this a domain- and application-specific *ontology* is needed.

Multilingual IE in the ZENON system

The ZENON system was built to analyze English HUMINT reports. Version 2 will extend the original system so that the result of analyzing English and Dari text can be processed together. In Figure 1 the main idea of our *multilingual information extraction system ZENON version 2*

[2] Since the KFOR corpus is classified, it is not freely available.

is shown. In the upper half of the illustration the processing of the English documents is shown. As already described above the named entities and the action types are determined by chunk-parsing using Gazetteers and handcrafted language-specific grammars. Named entities and action types in each sentence of each document are then combined to realize the content analysis. The result of both processing steps is a formal representation of actions, entities and content. This output from various documents can be combined and filtered according to analysis needs. This result can be used by the analyst - directly or in graphical form. The concepts of the application domain are formally described in the application-specific ontology, which is used during the processing steps.

In the lower part of Figure 1 the processing of the Dari documents is illustrated. For the identification of the named entities and the action types Gazetteers and handcrafted language-specific grammars have to be built. The result must be combined to form the content analysis. The formal description of the content analysis, the named entities and the action types can now be combined together with the formal result from the analysis of the English documents. In this fusion and filtering step XSLT is used to put together the content of documents written in different languages. Basis of putting together formal content is the use of the same ontology. Named entities and action types from different languages can be put together only if they are mapped to the same concepts during the information extraction process.

In the Dari part of the system a Dari-German dictionary will be integrated. The result of this part is then not only the formal content representation but also a word-to-word translation of the Dari document. This raw translation is possibly used by the analyst after analyzing the formal result or the graphical version of it. This might be the case if he needs even more information about the content of the document. The word-to-word translation will give this information and the analyst can decide whether a high-quality translation from a human translator is needed.

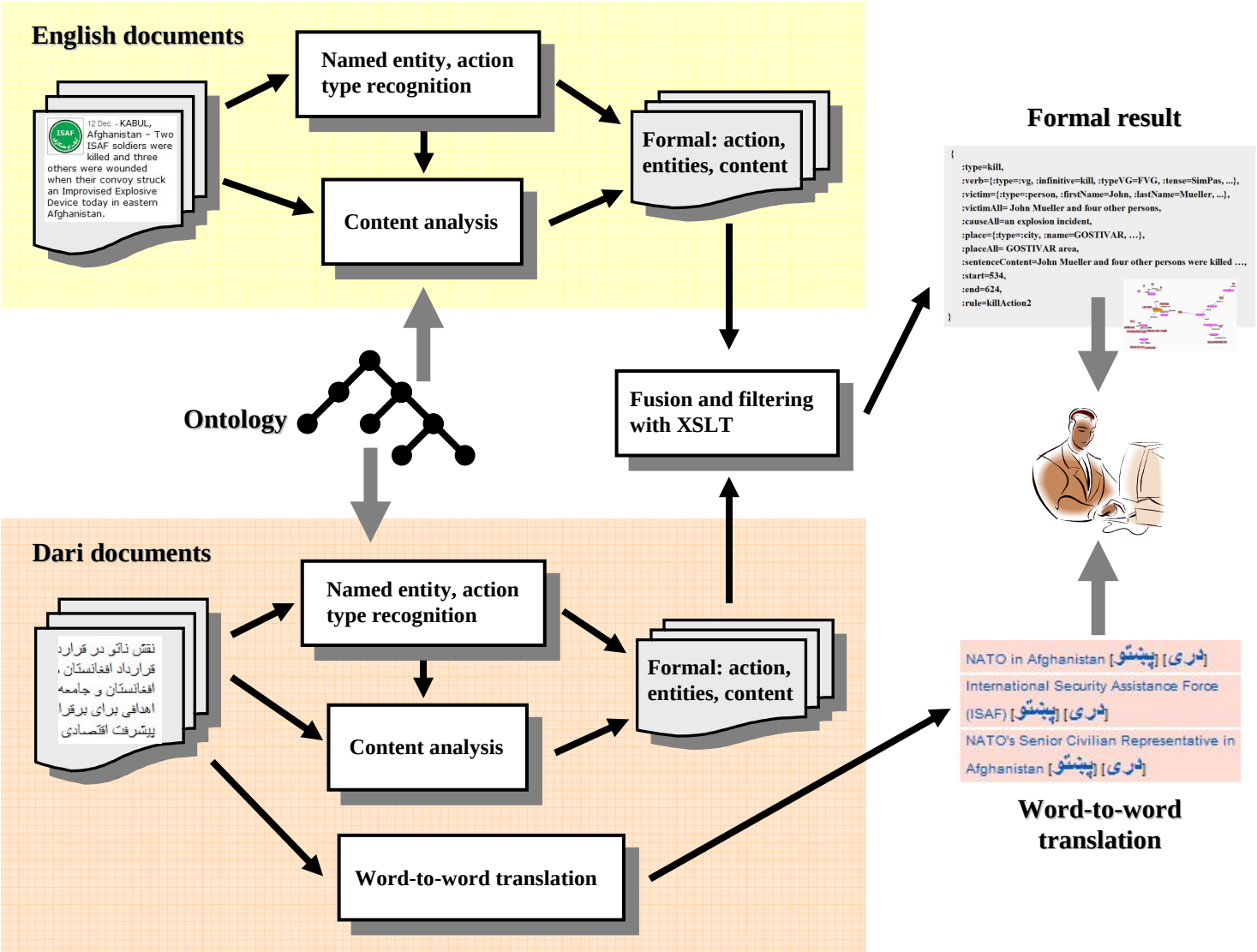

Figure 1: Main idea of multilingual information extraction

Dari processing resources and named entity recognition

We are in the process of realizing the idea of multilingual information extraction in the ZENON system version 2. The system will be able to process documents in English and in Dari. The functionality for the English language is those from version 1 of the ZENON system. For the Dari language natural language processing functionality was realized in a master thesis (cf. [Schwerdt, 2007a]). GATE components were built to recognize named entities, a Dari-German dictionary was integrated to get shallow word-to-word translation and a component was constructed to analyze the morphology of Dari verbs.

The built Dari GATE application contains the following natural language components: Dari Document Reset, Dari Tokenizer, Dari Sentence Splitter, Dari Gazetteer, Dari Grammar, Dari VP Chunker, and Dari Orthomatcher. The *Dari Document Reset* is used for resetting internal data structures at the beginning of the processing of a new Dari document. The *Dari Sentence Splitter* is the one delivered by the GATE system. From the Arabic components of GATE the Dari Tokenizer and Dari Grammar were taken and expanded. The Dari Gazetteer was created from scratch and contains the following lists of names in Dari (number of entries in each list): *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). For each of these lists rules for recognizing the equivalent named entity type was created. Figure 2 shows the result of the application of the *Province*, *FirstName* and *Number* rules.

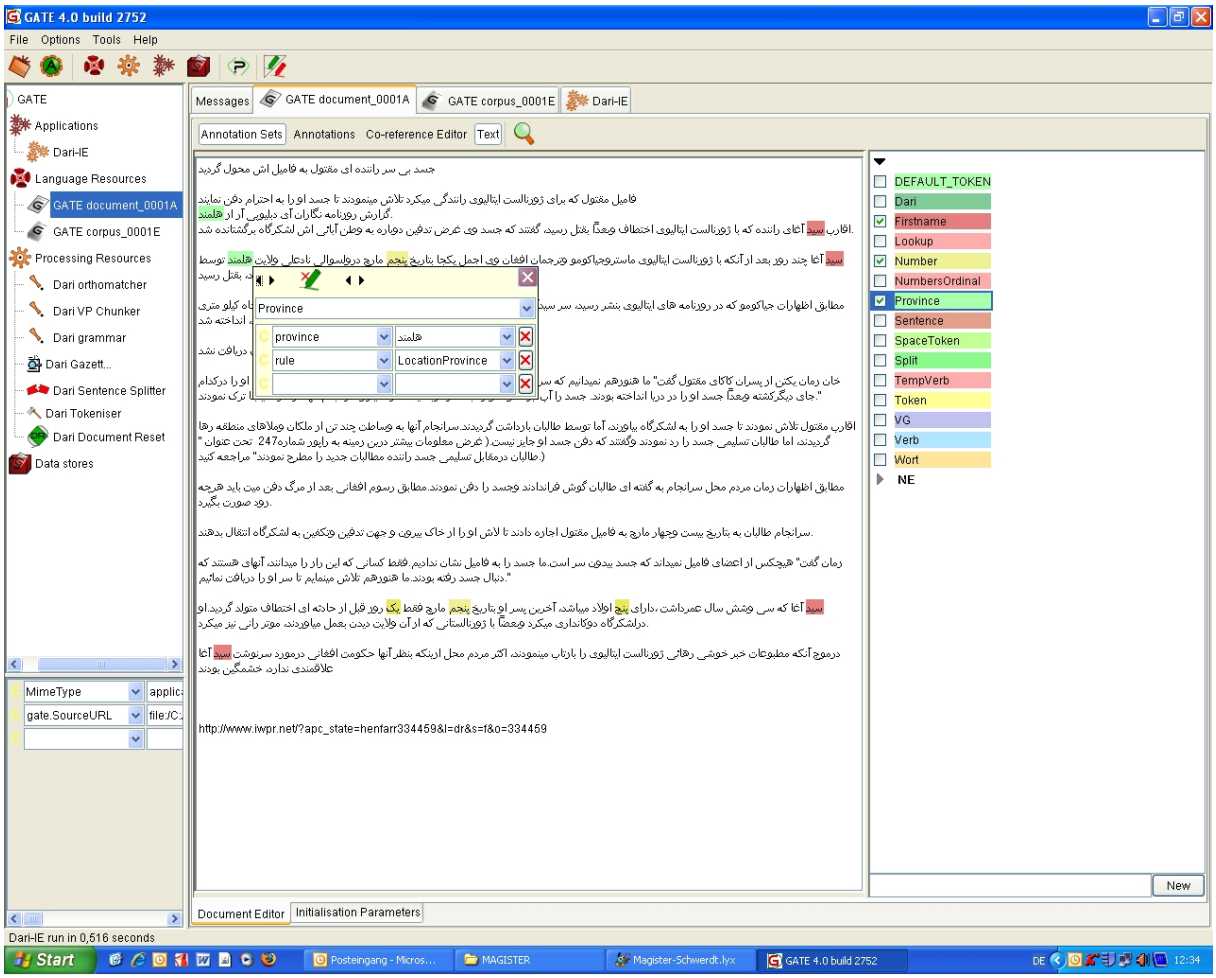

Figure 2: Named entity recognition

Dari-German dictionary

In the Dari application 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
- Herkunft (source of the entry)
- Sachgebiet (subject area)
- Deutsche Übersetzung (German translation)
- Deutsche Wortart (German part of speech)
- Deutscher Genus (German grammatical gender)
- Anzahl der deutschen Wörter (Number of words in German)
- Dari-Lemma (Dari lemma/headword)
- Dari-Numerus (Dari number)
- Englische Übersetzung (English translation)
- Quelle (source of the Dari word)
- Dari-Wortart (Dari part of speech)
- Dari Wortanzahl (Number of words in Dari)
- Dari-Dialekt (Dari dialect)
- Beschreibung (description)
- Häufigkeit (frequency)

In Figure 3 two dictionary entries are shown. Note that a feature is only shown if the feature value is not empty.

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

Figure 3: Two dictionary entries

During the processing a word-to-word translation is produced and the translation result is stored in the annotation type *Wort* (s. Figure 4).

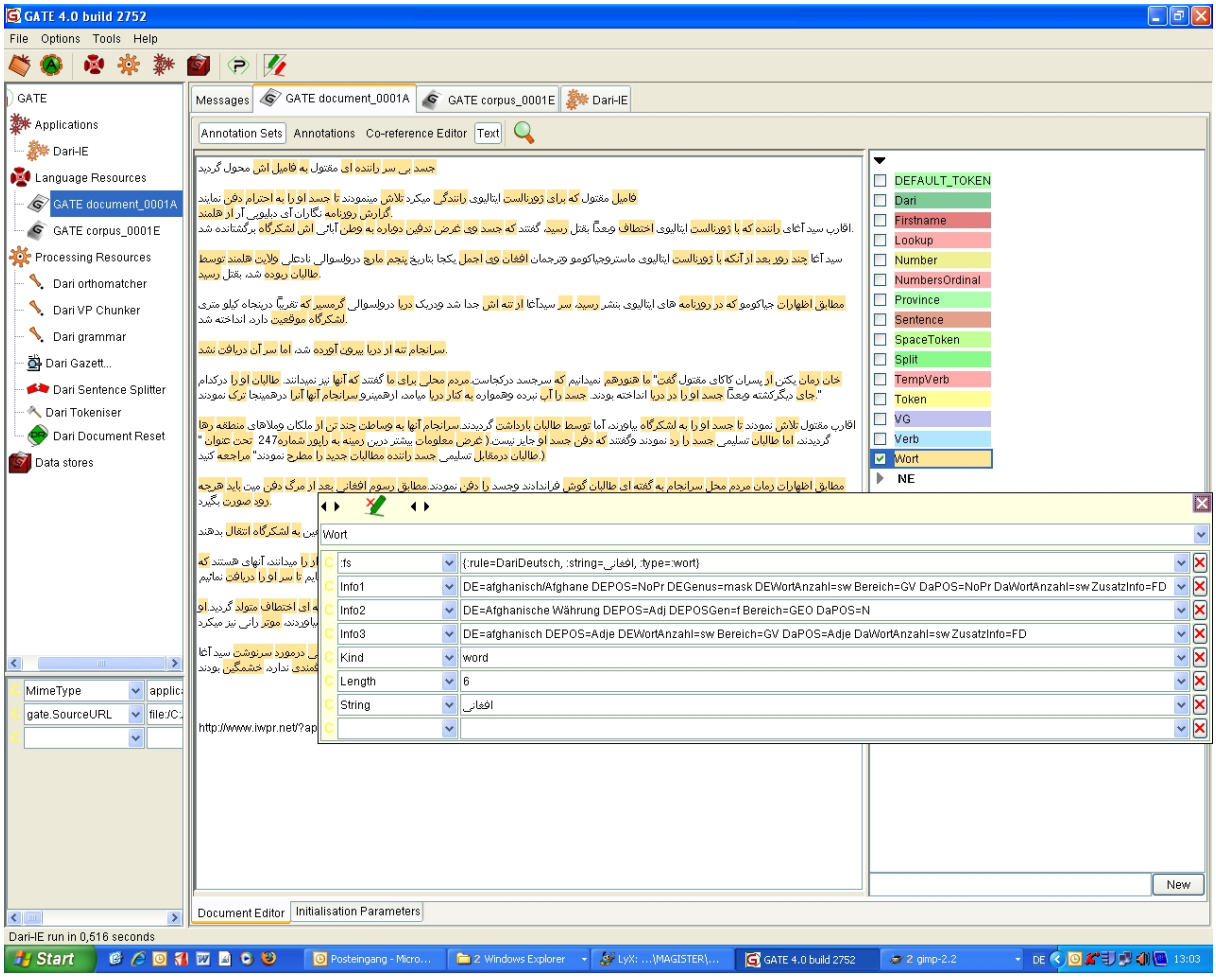

Figure 4: Annotation *Wort* with different word meanings

Morphological analysis of verb groups

For the identification of the action types used in the Dari documents the verbal phrases must be analyzed. For this purpose a morphological analysis was realized for the most important verbs in Dari. Two full-form lexica were created. The first one contains 39,552 inflectional past tense forms of the 8,818 verbs from the Dari-German dictionary. In Figure 5 the infinitive *budan* ("to be"), the stem *bud* and the inflectional past tense forms *budam*, …, *budand* are shown.

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

Figure 5: The infinitive *budan* ("to be")

The second full-form lexicon contains 6,592 past-participle forms of the verbs.

Due to the fact, that the forms of the present tense are irregular the morphological analysis of the present tense is restricted – at the moment - to the 84 most often used verbs (but 8,818 verbs are detected if they are in past tense). The analysis provides information about *stem*, *mood* (indicative, subjunctive, and imperative), *number*, *tense*, *negation*, *voice* and *translation of the infinitive*. Compound tenses are also handled. In Figure 6 the result of the analysis of single verbs is shown. In Figure 7 the result for compound verb phrases is given.

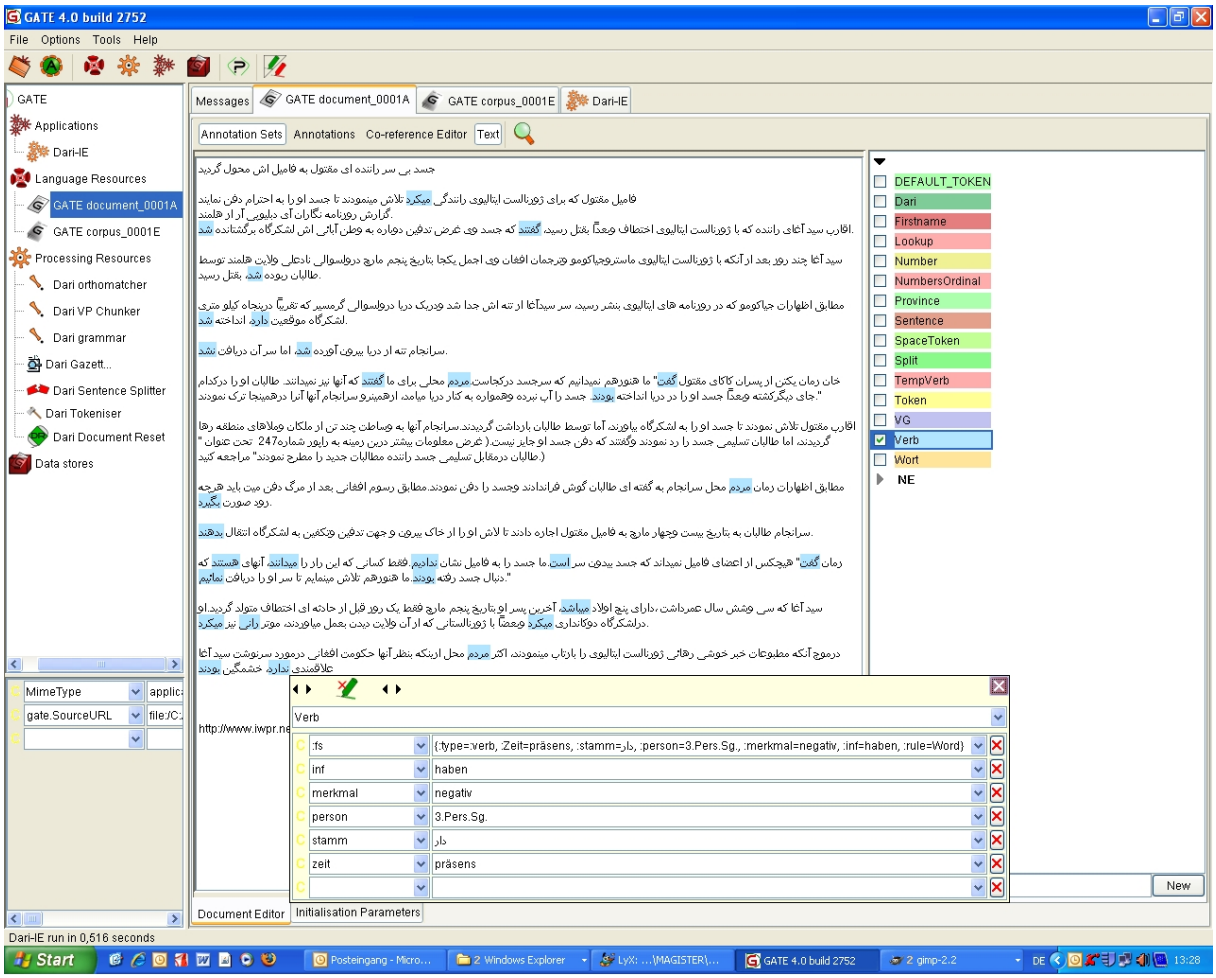

Figure 6: Verb annotation

The identified stems are associated with the same action types used during the analysis of the English verbal phrases.

We are in the process of integrating the modules for the semantic analysis of English and Dari documents.

Open problems

There are of course a lot of possibilities to improve the implementation of the multilingual information extraction system. One important issue is the transliteration of Dari names to those written in roman scripts. The problem is extended due to spelling variants in both languages. For this problem, we use at the moment only an ad hoc solution. A more principal solution is needed in which for each name *one* canonical form and its spelling variants are defined.

Another problem is the integration of the word-to-word translation with the information extraction. At the moment, the raw translation is not integrated into the formal description of the extracted information.

A more general problem is to get the same coverage of linguistic data (dictionaries, grammars …) and functionality (POS tagger, morphology …) for less commonly spoken languages.

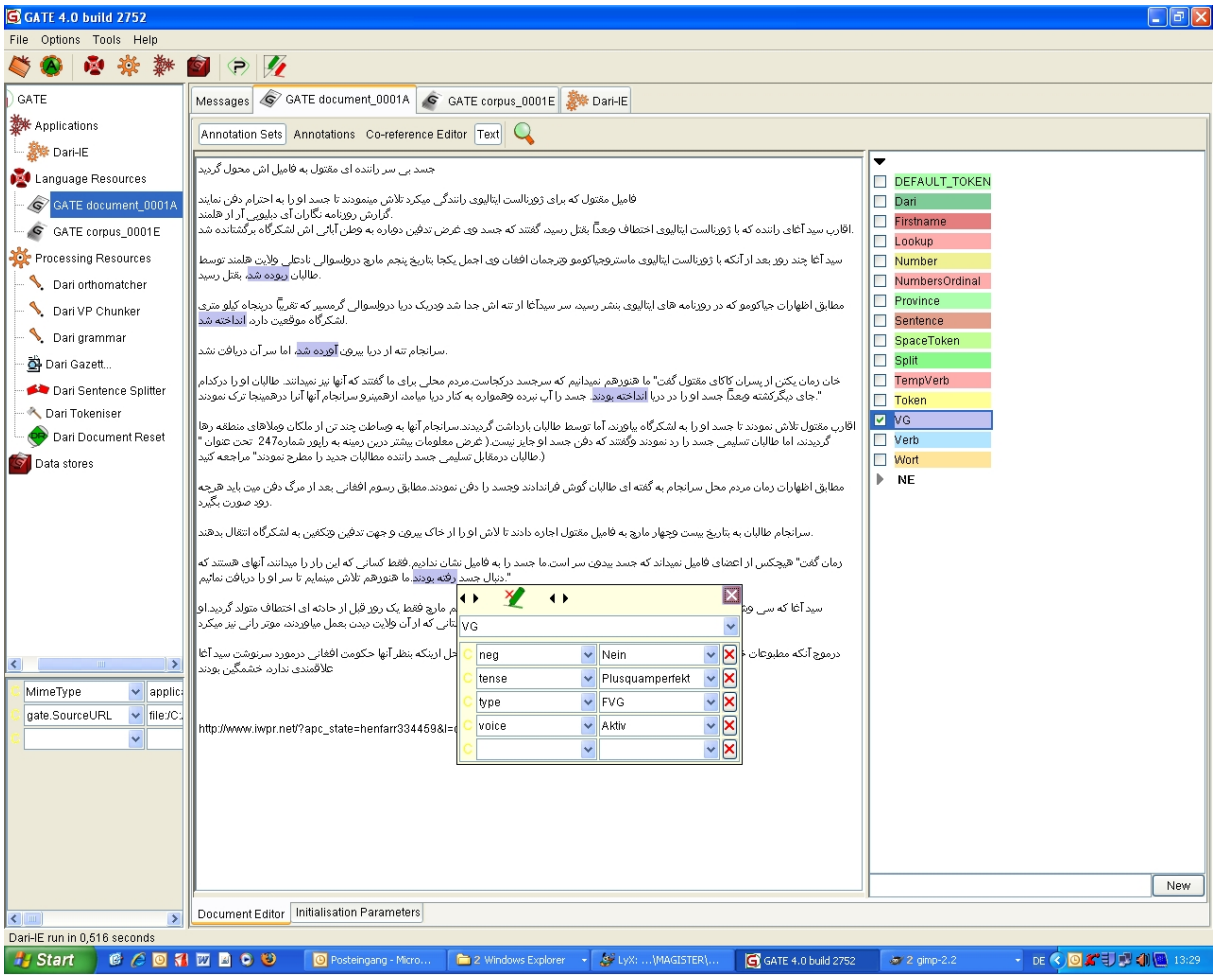

Figure 7: Compound verb phrase

4. Conclusion

In this paper, we have presented how we extended the ZENON system to do information extraction from texts written in different languages - in our approach for English and Dari. We presented the main idea of multilingual information extraction and showed how named entity recognition was realized for Dari texts. A Dari-German dictionary extends the capability of the information extraction functionality. This was presented, as well as the morphological analysis of Dari verb groups. Finally, open problems were discussed. At the moment, we are in the process of integrating the modules for the semantic analysis of English and Dari documents.

References

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

[Cunningham, 2002] Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V. *GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications*. Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02). Philadelphia, July 2002.

[Hecking, 2003a] Hecking, M. *Information Extraction from Battlefield Reports*. In: Proceedings of the 8th International Command and Control Research and Technology Symposium (ICCRTS), Washington, DC, U.S.A., 2003.

[Hecking, 2003b] Hecking, M. *Analysis of Free-form Battlefield Reports with Shallow Parsing Techniques*. Paper presented at the RTO IST Symposium on „Military Data and Information Fusion“, held in Prague, Czech Republic, October 20-22, 2003.

[Hecking, 2004a] Hecking, M. *Informationsextraktion aus militärischen Freitextmeldungen*. Forschungsgesellschaft für Angewandte Naturwissenschaft e.V. (FGAN), Wachtberg, Germany, FKIE-Bericht Nr. 74, 2004.

[Hecking, 2004b] Hecking, M. *How to Represent the Content of Free-form Battlefield Reports*. In: Proc. of the 2004 Command and Control Research and Technology Symposium (CCRTS) "The Power of Information Age Concepts and Technologies", June 15-17, 2004, San Diego, California.

[Hecking, 2005a] Hecking, M. *Domänenspezifische Informationsextraktion am Beispiel militärischer Meldungen*. In: A.B. Cremers, R. Manthey, P. Martini, V. Steinhage (Hrsg.) "INFORMATIK 2005”, Band 2, Lecture Notes in Informatics, Volume P-68, Bonn, 2005.

[Hecking, 2006a] Hecking, M. *Content Analysis of HUMINT Reports*. In: Proc. of the 2006 Command and Control Research and Technology Symposium (CCRTS) "THE STATE OF THE ART AND THE STATE OF THE PRACTICE", June 20-22, 2006, San Diego, California.

[Hecking, 2006b] Hecking, M. *Navigation through the Meaning Space of HUMINT Reports*. In: "Proceedings of the 11th International Command and Control Research and Technolgy Symposium”, September 26-28, 2006, Cambridge, UK.

[Hecking, 2006c] Hecking, M. *Das KFOR-Korpus als Ergebnis semantisch annotierter militärischer Meldungen*. Forschungsgesellschaft für Angewandte Naturwissenschaft e.V. (FGAN), Wachtberg, Germany, FKIE-Bericht Nr. 124, 2006.

[Hecking, 2007a] Hecking, M. *The KFOR Text Corpus*. In: "Proceedings of the 12th International Command and Control Research and Technolgy Symposium”, June 19-21, 2007, Newport, U.S.A.

[NCW, 2001] Department of Defense. *Network Centric Warfare – Report to Congress*. 27 July 2001.

[Poibeau, 2007a] Poibeau, T. & Saggion, H. *Proceedings of the International Workshop "Multi-Source, Multilingual Information Extraction and Summarization"*, held in conjunction with RANLP-2007, September 27-28, 2007, Borovets, Bulgaria.

[Pouliquen, 2007a] Pouliquen, B. & Steinberger, R. & Belyaeva, J. *Multilingual multi-document continuously-updated social networks*. In: Proceedings of the International Workshop "Multi-Source, Multilingual Information Extraction and Summarization", held in conjunction with RANLP-2007, September 27-28, 2007, Borovets, Bulgaria.

[Schwerdt, 2007a] Schwerdt, C. *Analyse ausgewählter Verbalgruppen der Sprache Dari zur multilingualen Erweiterung des ZENON-Systems*. Forschungsgesellschaft für Angewandte Naturwissenschaft e.V. (FGAN), Wachtberg, Germany, FKIE-Bericht Nr. 146, 2007.

[Steeneken, 1996] Steeneken, H. J. M. *Potentials of Speech and Language Technology Systems for Military Use: an Application and Technology Oriented Survey*. NATO, Technical Report, AC/243(Panel 3)TP/21, 1996.