15th ICCRTS
"The Evolution of C2"

A Tajik Extension of the Multilingual Information Extraction System ZENON

Track 1: Concepts, Theory, and Policy

Dr. Matthias Hecking (point of contact)
Tatjana Sarmina – Baneviciene

Fraunhofer Institute for Communication,
Information Processing and Ergonomics FKIE
Neuenahrer Straße 20
53343 Wachtberg
Germany

Phone: +49 228 9435 576
Fax: +49 228 9435 685
matthias.hecking@fkie.fraunhofer.de

A Tajik Extension of the Multilingual Information Extraction System ZENON

Dr. Matthias Hecking,
Tatjana Sarmina – Baneviciene
Fraunhofer Institute for Communication,
Information Processing and Ergonomics FKIE
Neuenahrer Straße 20
53343 Wachtberg, Germany
matthias.hecking@fkie.fraunhofer.de

Abstract

The new deployments of the German Federal Armed Forces create the necessity to analyze large quantities of intelligence reports and other documents written in different languages. We set up the research project ZENON, in which a multilingual information extraction approach is used for (partial) content analysis from texts written in different languages. At the moment the ZENON system is able to (partially) process English documents and documents written in Dari that are restricted in structure and vocabulary. In this paper, we present the functionality to do information extraction for Tajik texts as well. The Tajik module created will further extend our research system. We show how named entities, verbs and compound verb phrases from documents written in Tajik are extracted and represented. We also show how a simple word-to-word translation is integrated into the system. After a short introduction, the current multilingual information extraction project ZENON is explained. In the main part of the paper, our approach to information extraction from Tajik texts is described in detail.

1. Introduction

The new deployments of the German Federal Armed Forces (Bundeswehr) cause the necessity to analyze large quantities of intelligence reports and other documents written in different languages. Especially the *content analysis* of free-form texts is important for any information operation. During the content analysis the *actions* described and *entities* involved are extracted from the texts, combined (fused) and stored for further processing. A *partial* content analysis can be created 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 a *multilingual IE* approach is used for the (partial) content analysis from texts written in different languages. The overall objective of this project is to demonstrate that it is possible to use state-of-the-art natural language processing techniques to extract and combine military relevant knowledge from free-form texts even for rare languages. An expected advantage of systems like ZENON is the increased productivity of the *intelligence analyst*. He might analyze and combine information from more intelligence reports and from more open sources than without such automatic support. Even information from texts written in languages the analyst does not understand is accessible.

At the moment the ZENON system is able to process *English* documents (similar in structure and vocabulary to HUMINT reports from the KFOR deployment of the Bundeswehr) and documents written in *Dari* (cf. [Hecking, 2009a], [Hecking, 2008a], [Hecking, 2007a],

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

[Hecking, 2006a], [Hecking, 2006b], [Hecking, 2005a], [Hecking, 2004a], [Hecking, 2004b], [Hecking, 2003a], [Hecking, 2003b]). The knowledge about the actions and named entities is identified from each sentence and the content of the sentences are represented formally. These formal representations can be combined (fused, enhanced with further knowledge) and presented in a *graphically navigatable Entity-Action-Network*.

In the current version of the ZENON system the information extraction results from two different languages (English and Dari) are combined. Beside the information extraction the system gives a simple word-to-word-translation for Dari (to German) to further support the analyst. This allows the analyst to access information from Dari texts without knowing these languages. The automatic processing of the texts also extends the volume of these texts the analyst can handle. In view of the limited capabilities of the available natural language processing techniques the ZENON system is only an assistance of the analyst.

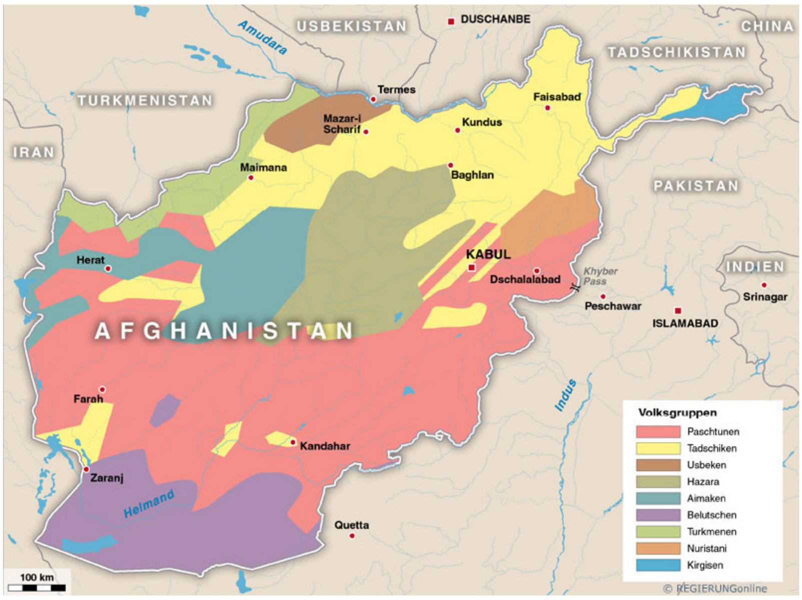

Figure 1: Ethnic groups in Afghanistan

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 first chose Dari for our multilingual approach because this language is spoken in the operational area of the German Bundeswehr. Also in the north (see Figure 1, yellow marking) the *Tajik* language (written in the Cyrillic alphabet) is used, also spoken by the Tajik people in Tajikistan. We have developed natural language processing functionality for Tajik and are about to expand the ZENON system with this Tajik module.

Due to missing annotated corpora we were not able to do comprehensive systematic evaluations for Dari and Tajik. For the English language we did some systematic evaluation for the named entity recognition using the *KFOR corpus*[2] (for some results cf. [Hecking, 2007a]).

This paper is structured as follows. First, a short introduction to the current ZENON system is given. In the main part of the paper, the Tajik extension of the ZENON system is described. For this, the general concept and the Tajik functionality created are presented. Finally, we discuss open problems.

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

2. The Multilingual ZENON System

Multilingual information extraction is an current research topic (cf. [Poibeau, 2007a]). 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. If information written in different languages can be (partially) extracted and fused automatically – without the use of a human translator – this would speed up the information gathering process. This would also be the case if the performance of the information extraction for the different languages is developed differently.

For the implementation of multilingual IE, processing resources (lexica, dictionaries, grammars, morphological analyzers etc.) are needed for each language. Because corpora are often missing for rare languages (e.g., Dari and Tajik), the application of machine learning methods are often not possible. Therefore, the classical rule-based approach applies. For this reason, the grammars in the ZENON research system were created by hand (grammar writing). Beside language-specific lexica, dictionaries, grammars, morphological analyzers etc. a shared set of concepts is necessary. For this a domain- and application-specific *ontology* is needed.

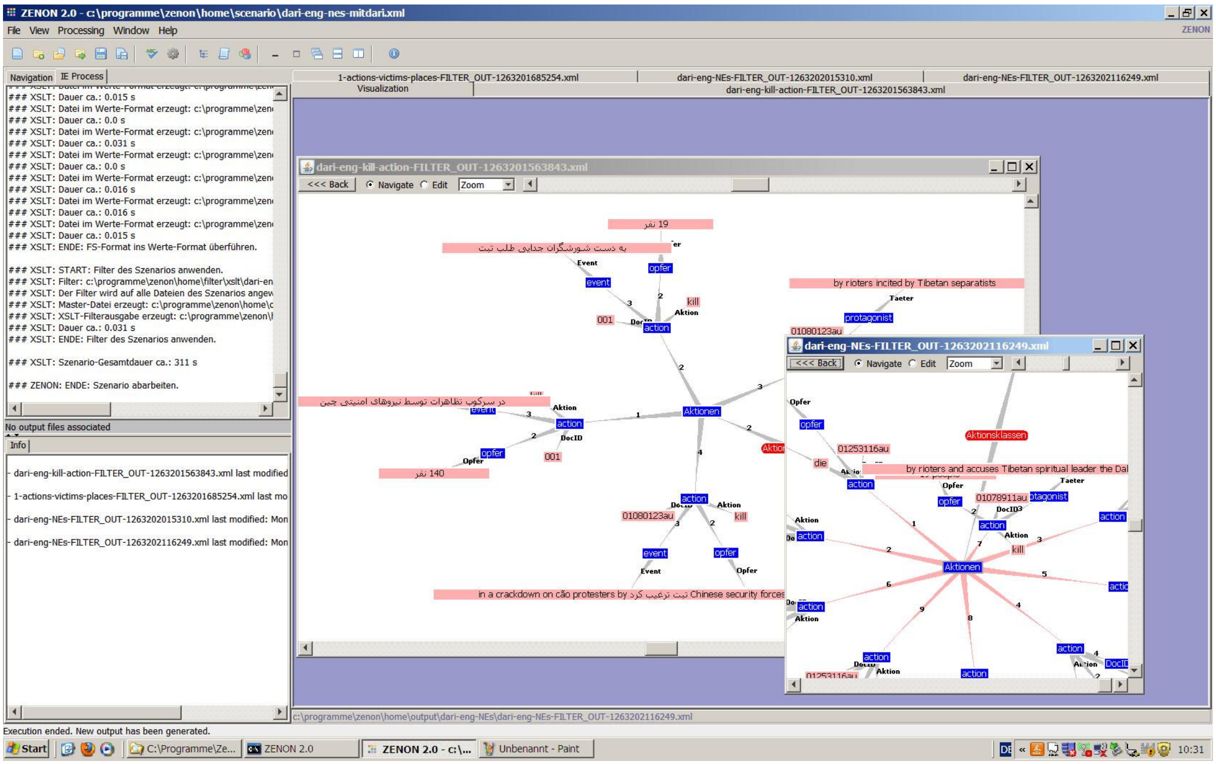

Figure 2: The ZENON system

In the current implementation the research system ZENON (see Figure 2) is able to do a (partial) content analysis of English and Dari documents. The English functionality was developed by using English HUMINT reports (cf. [Hecking, 2007a], [Hecking, 2006c]) from the KFOR deployment of the German Federal Armed Forces.

The content of these HUMINT 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 entities involved.

The ZENON system uses a *shallow syntactic approach* based on *chunk-parsing* (cf. [Hecking, 2004a]) and *transducer*. The approach is called ‘shallow’ because only those parts of a sentence are analyzed which are of interest for the application. The main advantage of this approach is its robustness when confronted with ungrammatical sentences. The disadvantage is that relevant information may possibly be missed. The transducers are handcrafted grammars processed as finite automata.

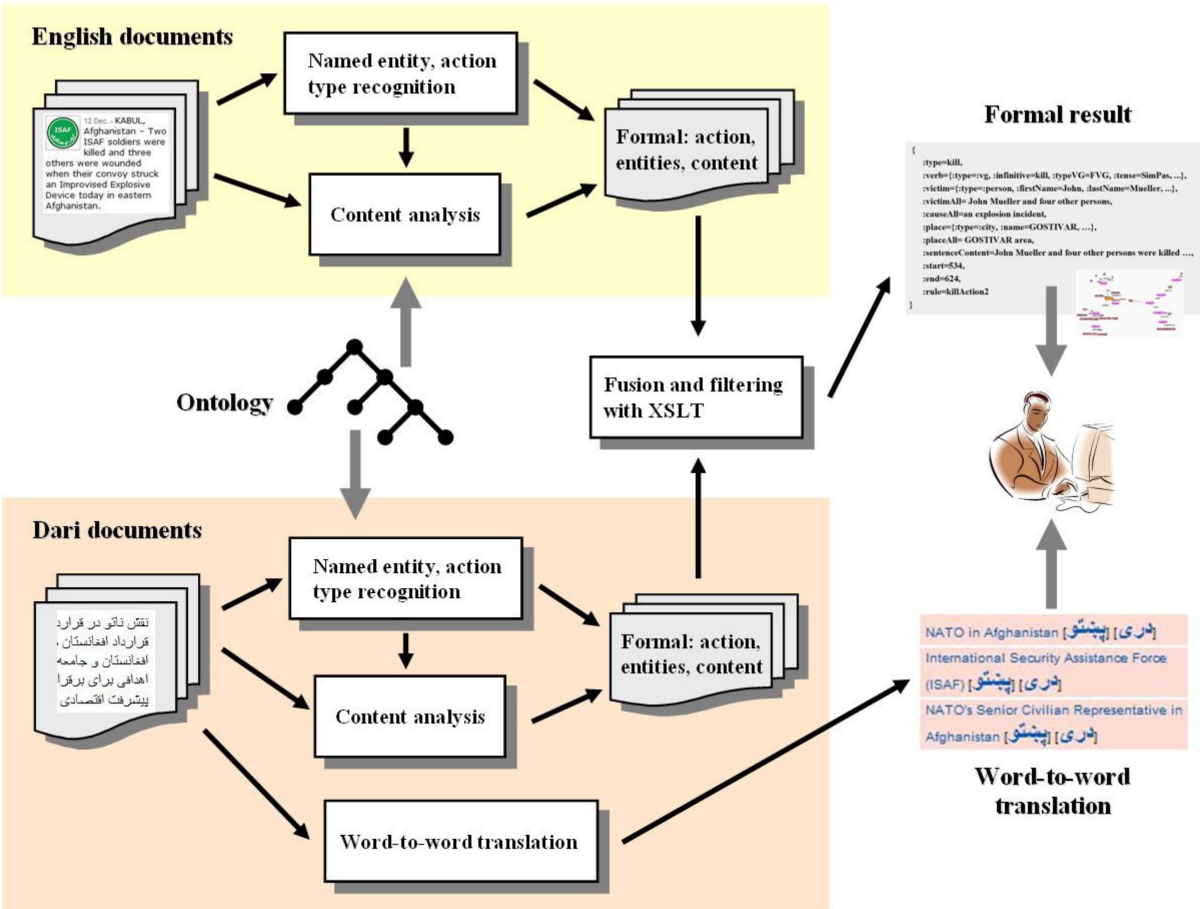

Figure 3: Architecture of the current ZENON system

During the processing of the English documents (see upper part in Figure 3) 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* (cf. [Hecking, 2004b]). 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 content extracted of different reports are combined and selected according to predefined XSLT (Extensible Stylesheet Language Transformation) sheets. These sheets implement the information needs of the analyst. Then, the result of the analysis is depicted graphically and can be navigated interactively (see the graphs in Figure 2).

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 English named entities were developed: *City, Company, Coordinates, Country, CountryAdj, Currency, Date, GeneralOrg, MilitaryOrg, Number, Percent, Person, PoliticalOrg, Province, Region, River, Time* and *Title*. For the verb phrases analysis the ZENON prototype uses various transducers (more than 100) to recognize finite and non-finite verbal phrases, modal verb phrases, participles and special composed verb expressions.

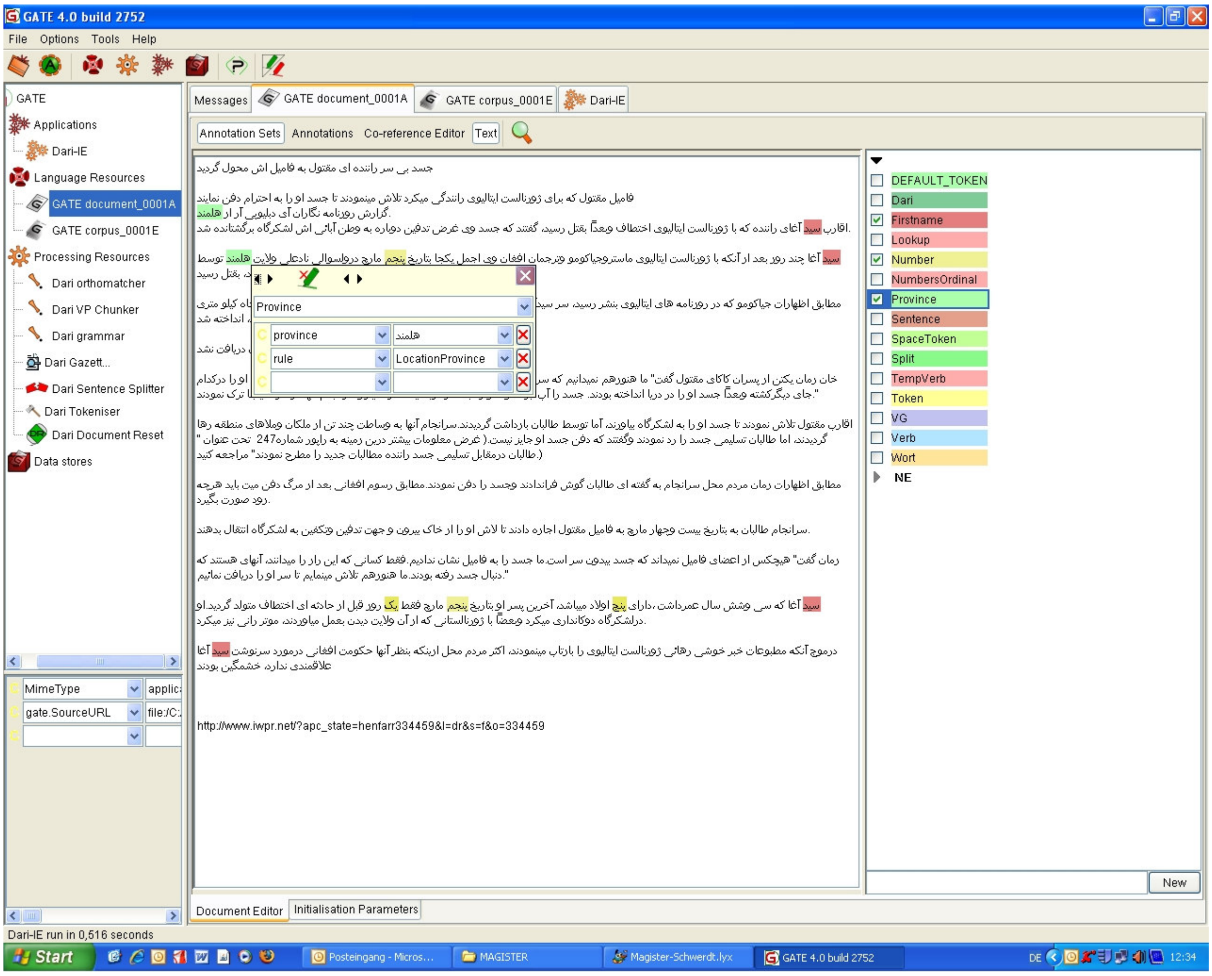

Figure 4: Named entity recognition in Dari

For the construction of the ZENON system (for both languages) the toolbox GATE (General Architecture for Text Engineering, cf. [Cunningham, 2002]) is used. 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 and/or extended to implement the natural language processing parts of the ZENON research system (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. Grammars are written with the JAPE (Java Annotation Pattern Engine) formalism.

In the current implementation the ZENON system is also able to process Dari documents (see the lower part in Figure 3). The Dari module contains various natural language components (cf. [Hecking, 2008a], [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.

In the Dari module the submodule *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 (more then 1,800 names): *city_afgh*, *city_world*, *country_world*, *days*, *female_names*, *male_names*, *months*, *numbers*, *ordinals*, *province_afgh*. For each of these lists rules for recognizing the equivalent named entity type are available. Figure 4 shows the recognition results for the named entity types 'Province', 'FirstName' and 'Number'.

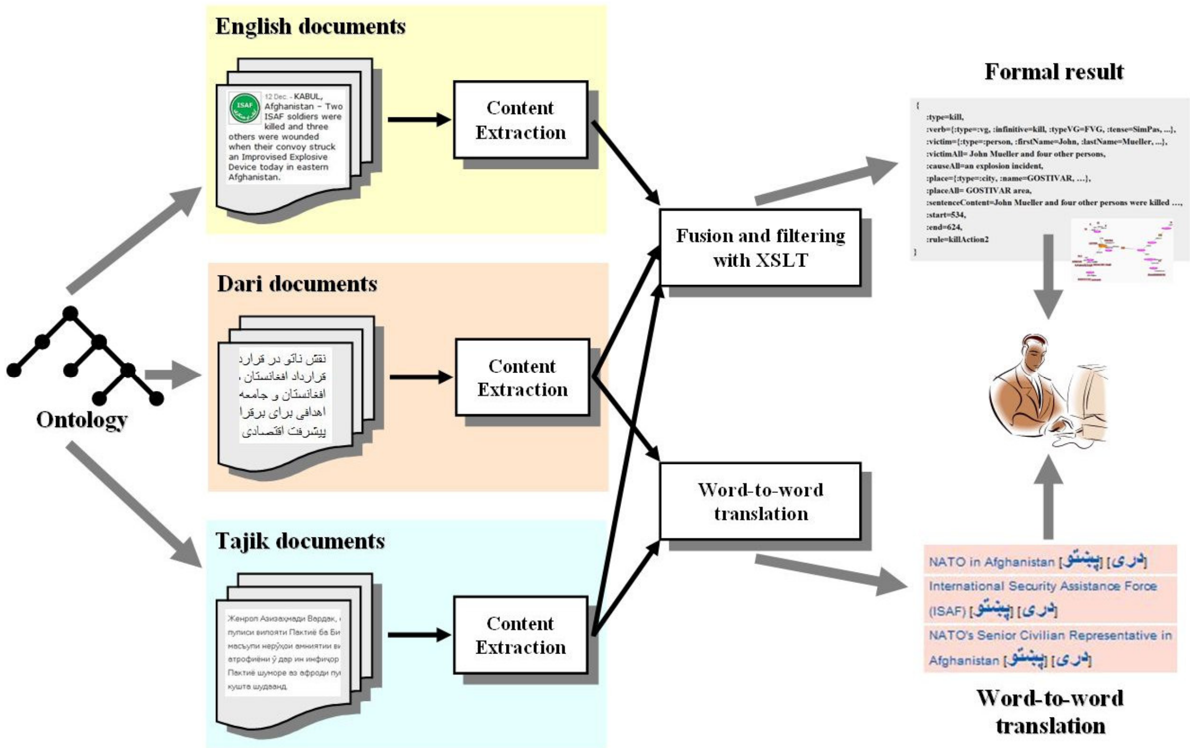

Figure 5: The multilingual ZENON with the Tajik module

In the Dari module a *Dari-German full-form dictionary* with approximately 48,000 entries is integrated. So the result of the module is not only the formal content representation but also a word-to-word translation (see Figure 3) of the Dari document. This raw translation can be used by the analyst to decide whether a high-quality translation from a human translator is needed. At the moment the translations for the words are stored in the annotation type *Wort*. A more user-friendly interface for the analyst has to be implemented yet.

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 created for the most often used verbs in Dari. Two full-form lexica were implemented. The first one contains 39,552 inflectional past tense forms of the 8,818 verbs from the Dari-German dictionary. The second full-form lexicon contains 6,592 past-participle forms of the verbs.

The formal description of the content analysis, the named entities and the action types from the Dari documents 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 content extraction process.

3. The Multilingual Tajik Extension of the ZENON System

The current implementation of the ZENON research system will be extended by natural language processing functionality for *Tajik* texts. The GATE module was already implemented (cf. [Sarmina, 2009a]), but it is not yet integrated into the overall system. After integrating the module the overall architecture will be like that shown in Figure 5. In addition to the English and Dari modules a Tajik module (lower part in the figure) analyses Tajik texts and creates also formal descriptions of named entities and verbs. These descriptions can be combined with the results of the English and Dari analysis. So with this approach it is possible to fuse information from sources written in the three different languages. The Tajik module also presents a word-to-word translation for the analyst (like the Dari module).

The Tajik language (Tajik Persian, тоҷикӣ, تاجیکی) is a modern version of Persian spoken in Central Asia. Most speakers of Tajik live in Tajikistan and Uzbekistan (see Figure 1, north of Afghanistan). Tajik is the official language of Tajikistan and a member of the Indo-European language family. The word order of Tajik is Subject-Object-Verb. The Tajik Persian grammar is almost identical to the classical Persian grammar (and the grammar of modern varieties such as Iranian Persian). Tajik is written in the Cyrillic alphabet (cf. [Tajik, 2010a]). For more information about the language, see [Rzehak, 1999a].

Documents available in Tajik with their English and Russian translations have been used for the development of the GATE components for the Tajik language. A broad collection of linguistically annotated Tajik texts is not currently available. The approach to build natural language components for Tajik using statistically-based machine learning methods is therefore not possible and the classical rule-based approach is used.

The *Tajik Tokenizer* and the *Tajik Sentence Splitter* are nearly identical to similar GATE components for the English language. The *Tajik Gazetteer* is used to identify names, as the basis for named entity recognition. In the following table the Gazetteer types with the number of entries are shown:

city_tg.lst (41)	mountain.lst (19)	person_full.lst (38)
city_world.lst (60)	name_all_male.lst (98)	pers_pron.lst (14)
country.lst (197)	name_all_female.lst (92)	person_relig.lst (11)
country_adj.lst (4)	name_tg_male.lst (526)	provinz.lst (21)
date_key.lst (12)	name_tg_female.lst (318)	region.lst (16)
day.lst (16)	number_letters.lst (156)	river.lst (40)
determiner.lst (8)	ordinal.lst (58)	sea.lst (17)
jobtitle.lst (98)	organization.lst (34)	time_unit.lst (5)
ministry.lst (15)	org_base.lst (96)	title.lst (20)
money.lst (9)	org_mil.lst (10)	title_female.lst (2)
month.lst (24)	org_terr.lst (53)	

All Gazetteer lists together contain about 2,100 entries. Included are also some Gazetteer lists with English names which may occur in the documents.

Based on the recognition of the Gazetteer list entries the named entities are identified with the help of the *Tajik NE Transducers*. These JAPE grammars generate new annotations that are used in subsequent processing steps. The following named entity types were implemented: *City, CommonOrg, Country, Date, GovernmentOrg, MilitaryOrg, Money, Person, Number, Gender, Jobtitle, Province, Region,* and *TerroristOrg*. All types have various features which are also determined during the recognition process. In Figure 6 the result for recognizing types 'Person' (dark red shaded) and 'Date' (pink shaded) are shown. In the small window in this figure the features with their values for a 'Person' named entity are depicted. In addition to features 'firstname' and 'lastname' the features 'rule', 'rule1' and 'rule2' are listed. They are used for trace purposes.

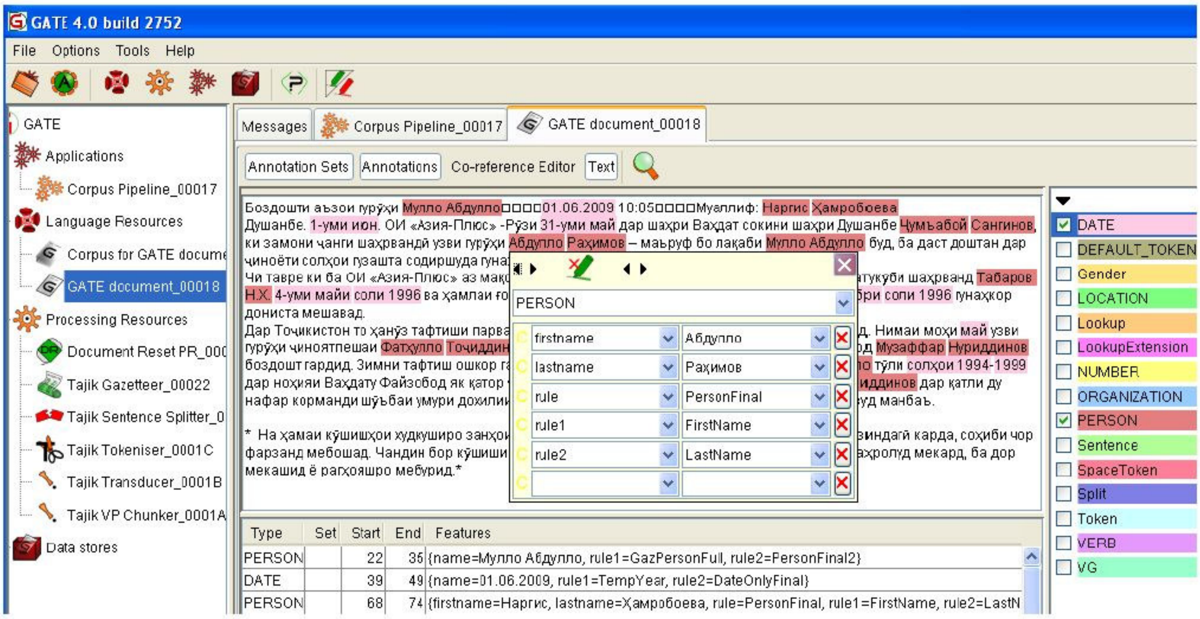

Figure 6: Named entity recognition of type 'Person' and 'Date'

The verbal phrases must be analyzed as a basis for the identification of the action used in the Tajik documents. For this purpose a (partial) morphological analysis was implemented. Four full-form lexica for Tajik verbs were created. They represent the *present-participle forms*, *past-participle forms*, *past-participle forms for compound verb phrases* and *compound verb phrases forms in past tense*.

The *Tajik VP-Chunker* implements through JAPE rules the analysis of finite (present, past tense, perfect, past perfect) and non-finite verb phrases, participles, adverbs and negations. The annotation created includes various syntactic information (e.g., tense, stem). To identify the number, infinitive and the verb stem the approach of *word stemming* was used. This is a simple method to identify the stem of irregular verb forms by pruning the letters from right to left till the stem is discovered.

The result of the analysis of simple verbs is stored in the annotation type *Verb*. In Figure 7 the result of the analysis of single verbs is shown. Different features contain the recognized information:

- infinitive ("Infinitiv"),
- mood ("Merkmal")
- person ("Person"),
- stem ("Stamm"),

- translation of the infinitive (“TranslationDE”),
- tense (“Zeit”),
- rule,
- string.

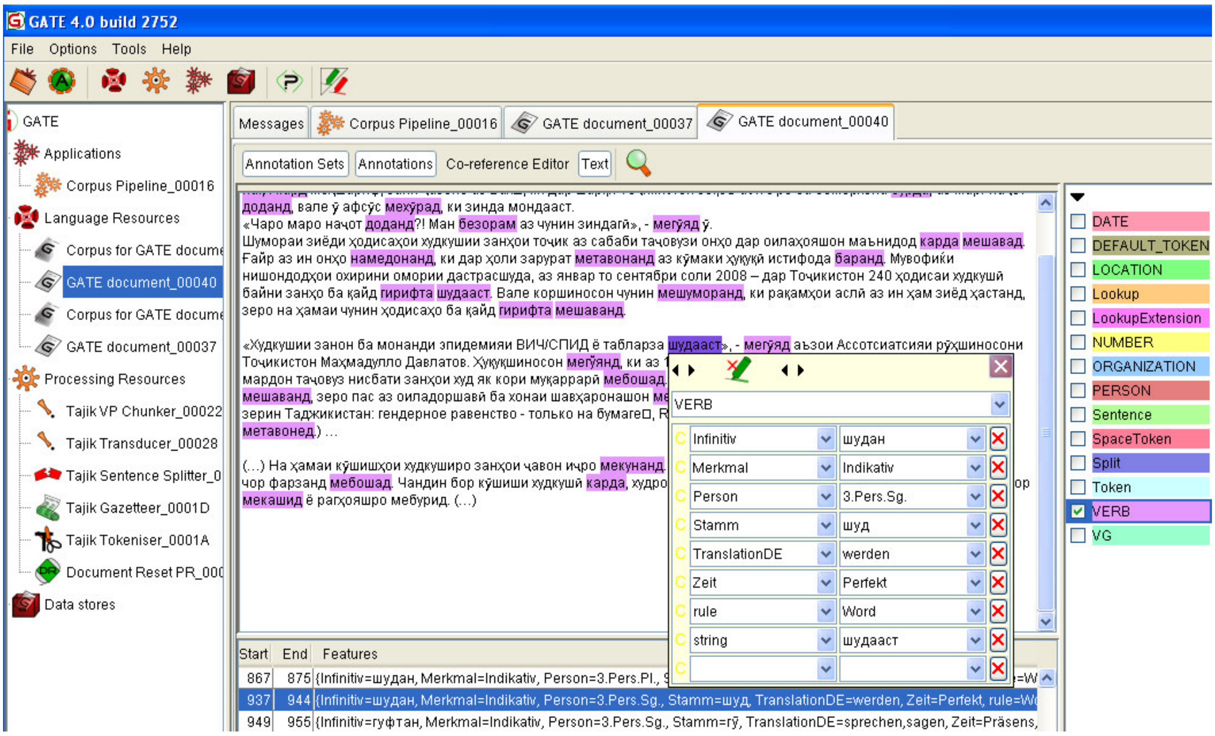

Figure 7: ‘Verb’ annotations

Compound verb phrases are also parsed. The result of this analysis is stored in the annotation type *VG* (verb group). Figure 8 shows the analysis results for some compound verb phrases. In the small window the features of the selected verb phrase are shown: the phrase is identified as a finite verb group (“FVG”) in present tense (“Präsens”) passive voice (“Passiv”).

The Tajik module contains also a translation submodule to give a *raw word-to-word translation* into German. This is an additional support for the analyst to decide whether a high-quality translation from a human translator should be created.

No online Tajik-German dictionary was freely available. Therefore a simple dictionary with 1,300 entries was created manually. For each entry the following information is available:

- Tajik lemma,
- Tajik part-of-speech (POS),
- Number of words in Tajik (TgWortanzahl) with values ‘sw’ (single word) or ‘mwX’ (multiple words with X words),
- German translation (TranslationDE).

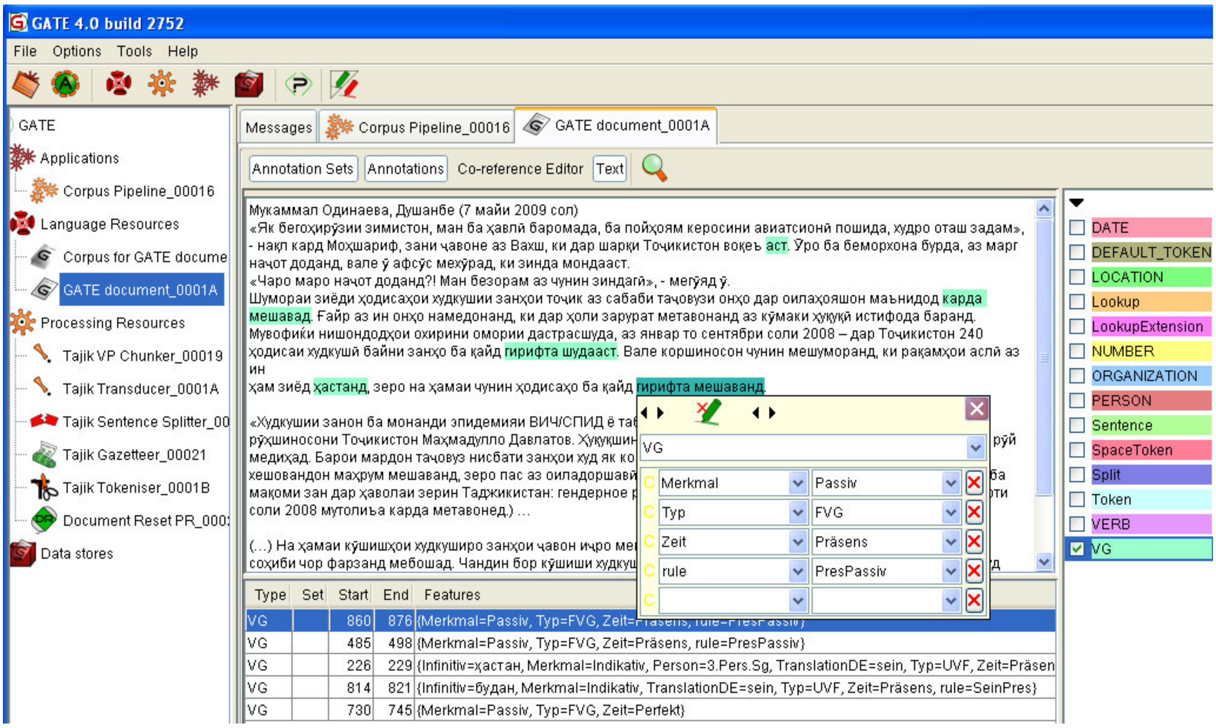

Figure 8: Annotations of compound verb phrases

For the part-of-speech an adapted version of the tag set used in the Hepple tagger was used (cf. [Cunningham, 2009a]). In Figure 9 some dictionary entries are shown ('NN' means noun, 'NNP' means proper noun singular).

Арабистони Саъудӣ:POS=NNP:TgWortanzahl=mw2:TranslationDE=Saudi Arabien
Артиллерист:POS=NN:TgWortanzahl=sw:TranslationDE=Artillerist
Аскари савора:POS=NN:TgWortanzahl=mw2:TranslationDE=Kavallerist
Аскар:POS=NN:TgWortanzahl=sw:TranslationDE=Soldat
Астронавт:POS=NN:TgWortanzahl=sw:TranslationDE=Astronaut
Астроном:POS=NN:TgWortanzahl=sw:TranslationDE=Astronom
Афгонистон:POS=NNP:TgWortanzahl=sw:TranslationDE=Afganistan
Афсар:POS=NN:TgWortanzahl=sw:TranslationDE=Offizier
Афғонистон:POS=NNP:TgWortanzahl=sw:TranslationDE=Afganistan
Ашт:POS=NNP:TgWortanzahl=sw:TranslationDE=Ascht
Балҷувон:POS=NNP:TgWortanzahl=sw:TranslationDE=Baldschuvon
Баҳри:POS=NN:TgWortanzahl=sw:TranslationDE=Meer
Баҳрнавард:POS=NN:TgWortanzahl=sw:TranslationDE=Matrose

Figure 9: Dictionary entries

The dictionary is implemented as a Gazetteer list. Figure 10 shows the lookup annotations for those words for which translations are available. The features of these annotations are then transferred into other annotations, e.g., into the Verb annotations (see in Figure 7 the Feature 'TranslationDE').

We are in the process of integrating the Tajik module into the current ZENON system.

There are of course a lot of possibilities to improve the implementation of the Tajik module. Most limitations are due to resource restriction. A larger dictionary would be an advantage. But the costs for the manual construction of it are beyond the scope of the ZENON project.

Not all possible grammatical constructions are processed. A nearly full coverage of the grammatical phenomena of Tajik would require much more time and effort.

At the moment the Tajik module recognizes named entities and verb phrases, but not action types and the combination of actions with their semantic roles (meaning of the whole sentence). After integrating the module into the ZENON system the overall system is therefore only able to process information from the three different languages based on named entities and verb phrases.

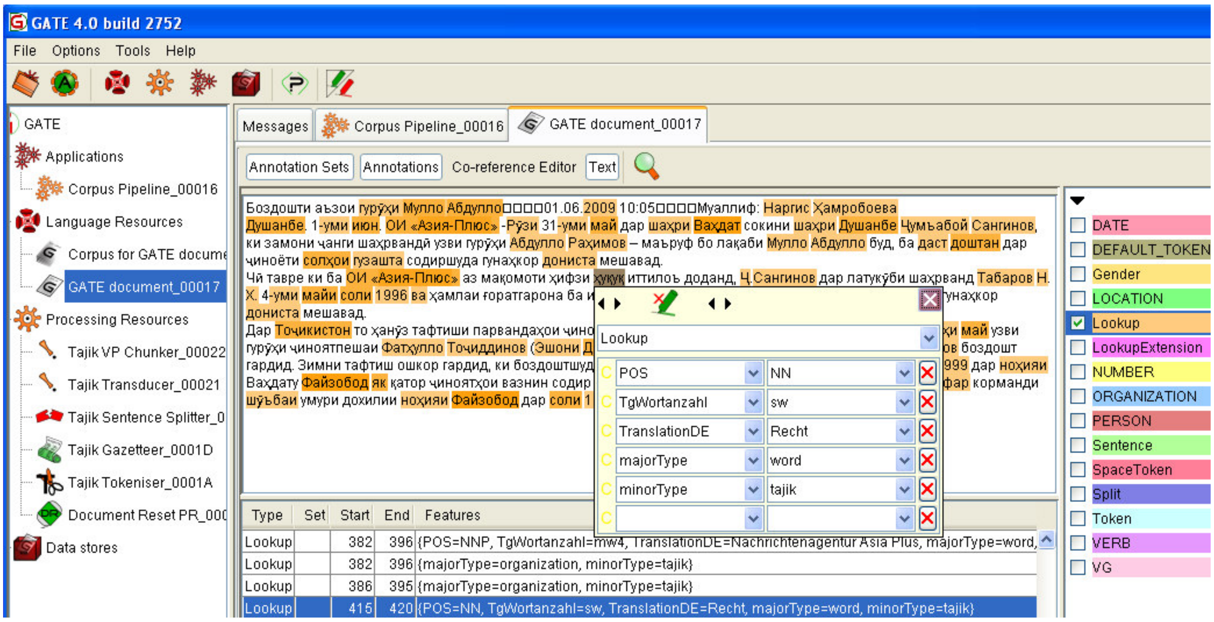

Figure 10: 'Lookup' annotations with word-to-word translations

Another limitation is the integration of the word-to-word translation. At the moment the translation is stored in feature values and not, e.g., inserted into the texts after the Tajik words in the Tajik document.

A more general problem is to get the same coverage of linguistic data (e.g., dictionaries, grammars) and functionality (e.g., POS tagger, morphology analyzer) for rare languages (like Dari and Tajik).

4. Conclusion

In this paper, we presented the functionality to perform information extraction for Tajik texts. The Tajik module created will be integrated into our research system ZENON. We show how named entities, verbs and compound verb phrases from documents written in Tajik are extracted and formalized. We also presented a simple Tajik-German word-to-word translation. Finally, open problems were discussed. At the moment, we are in the process of integrating the Tajik module into the current ZENON system.

We expect that systems like ZENON will increase productivity of the intelligence analyst. He might analyze and combine information even from texts written in languages the analyst does not understand.

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.

[Cunningham, 2009a] Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V., Ursu, C., Dimitrov, M., Dowman, M., Aswani, N., Roberts, I., Li, Y., Shafirin, A., & Funk, A. (2009). *Developing Language Processing Components with GATE Version 5 (a User Guide)*. The University of Sheffield, Sheffield, 2009.

[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.

[Hecking, 2008a] Hecking, M., Schwerdt, C., *Multilingual Information Extraction for Intelligence Purposes*. In: Proceedings of the 13th International Command and Control Research and Technolgy Symposium (ICCRTS), June 17-19, 2008, Bellevue, WA, U.S.A.

[Hecking, 2009a] Hecking, M., *Multilinguale Textinhaltserschließung auf militärischen Texten*. In: Verteilte Führungsinformationssysteme. Michael Wunder, Jürgen Grosche (Hrsg.), Springer-Verlag, 2009.

[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.

[Rzehak, 1999a] Rzehak, L. (1999). *Tadschikische Studiengrammatik*. Reichert Verlag, Wiesbaden, 1999.

[Sarmina, 2009a] Sarmina-Baneviciene, T., *Analyse spezifischer Probleme der tadschikischen Sprache zur multilingualen Erweiterung des ZENON-Systems*. Magisterarbeit, Philosophische Fakultät der Rheinischen Friedrich-Wilhelms-Universität zu Bonn, Bonn, Germany, 2009.

[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.

[Tajik, 2010a] *Wikipedia entry 'Tajik language'*. http://en.wikipedia.org/wiki/Tajik_language, (last accessed: 19.1.2010).