
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

1. Introduction
2. The Multilingual ZENON System
3. The Multilingual Tajik Extension of the ZENON System
4. Conclusion, References

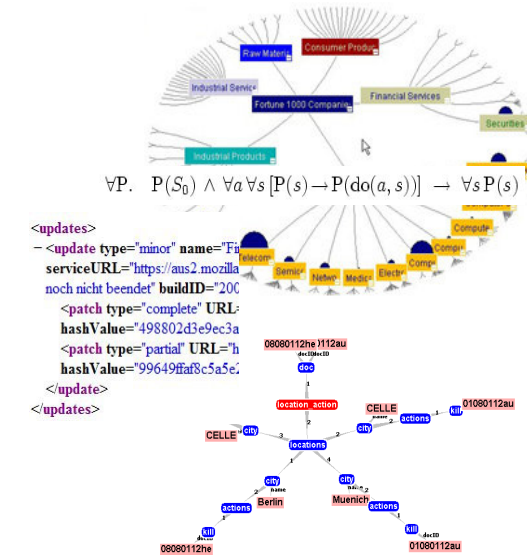
1. Introduction

Dr. M. Hecking



?

Extract knowledge

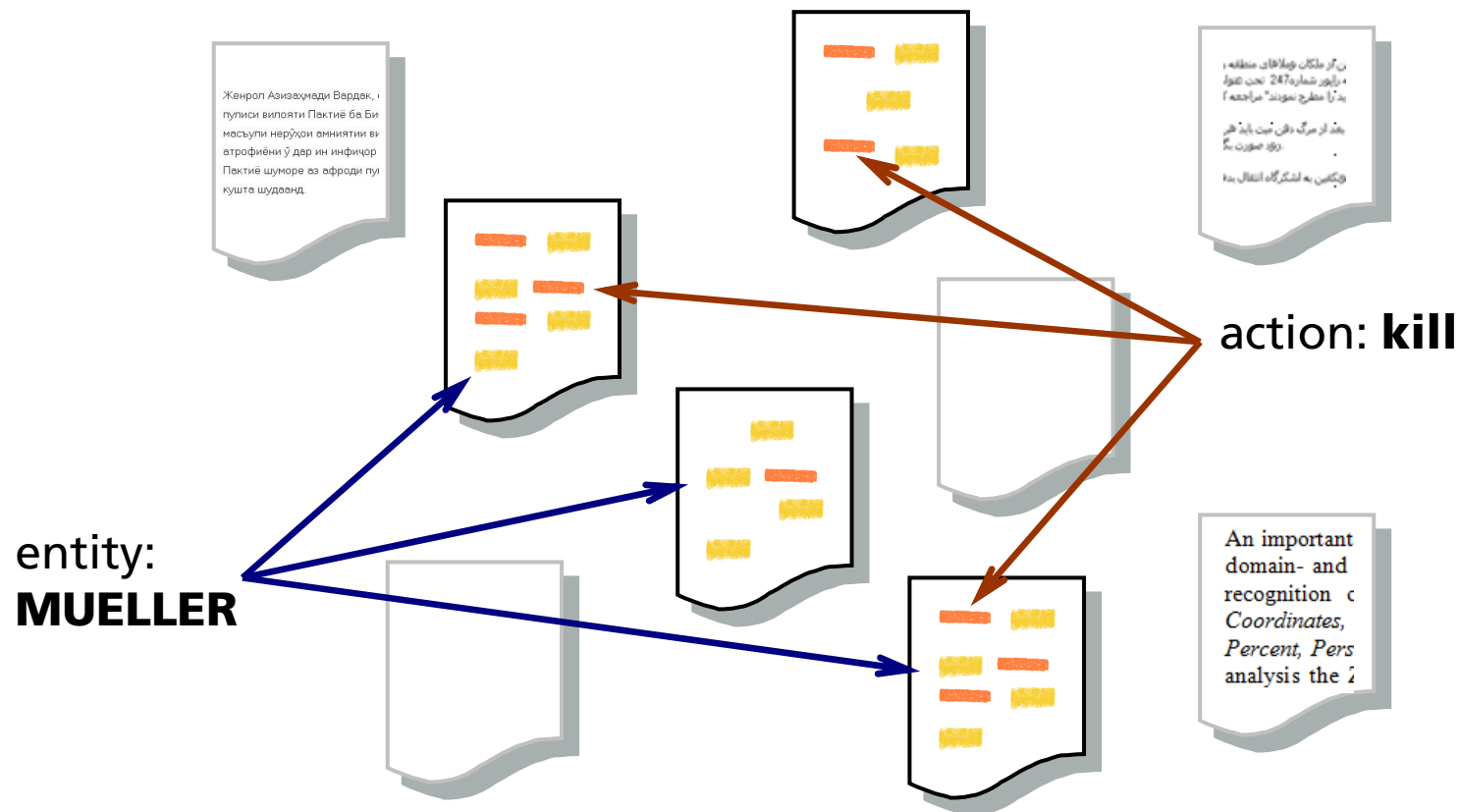


- Military relevant knowledge is available in different languages, formats and media.
- How to get the knowledge out of documents or audio files coded in different languages? How can we increase productivity of the intelligence analyst

2. The Multilingual ZENON System - I

Dr. M. Hecking

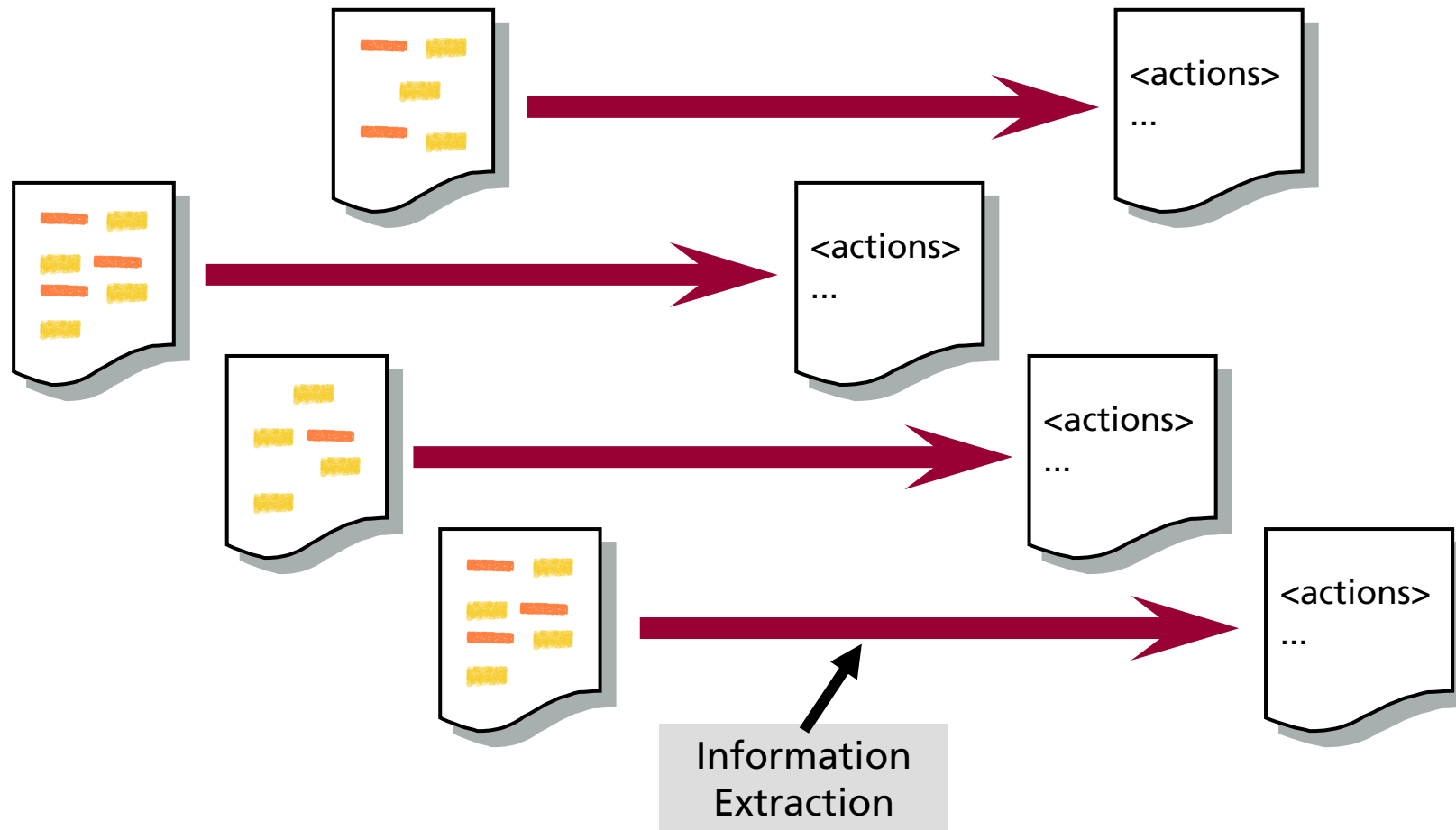
The knowledge about **actions** and **entities** is distributed over many reports.



2. The Multilingual ZENON System - II

Dr. M. Hecking

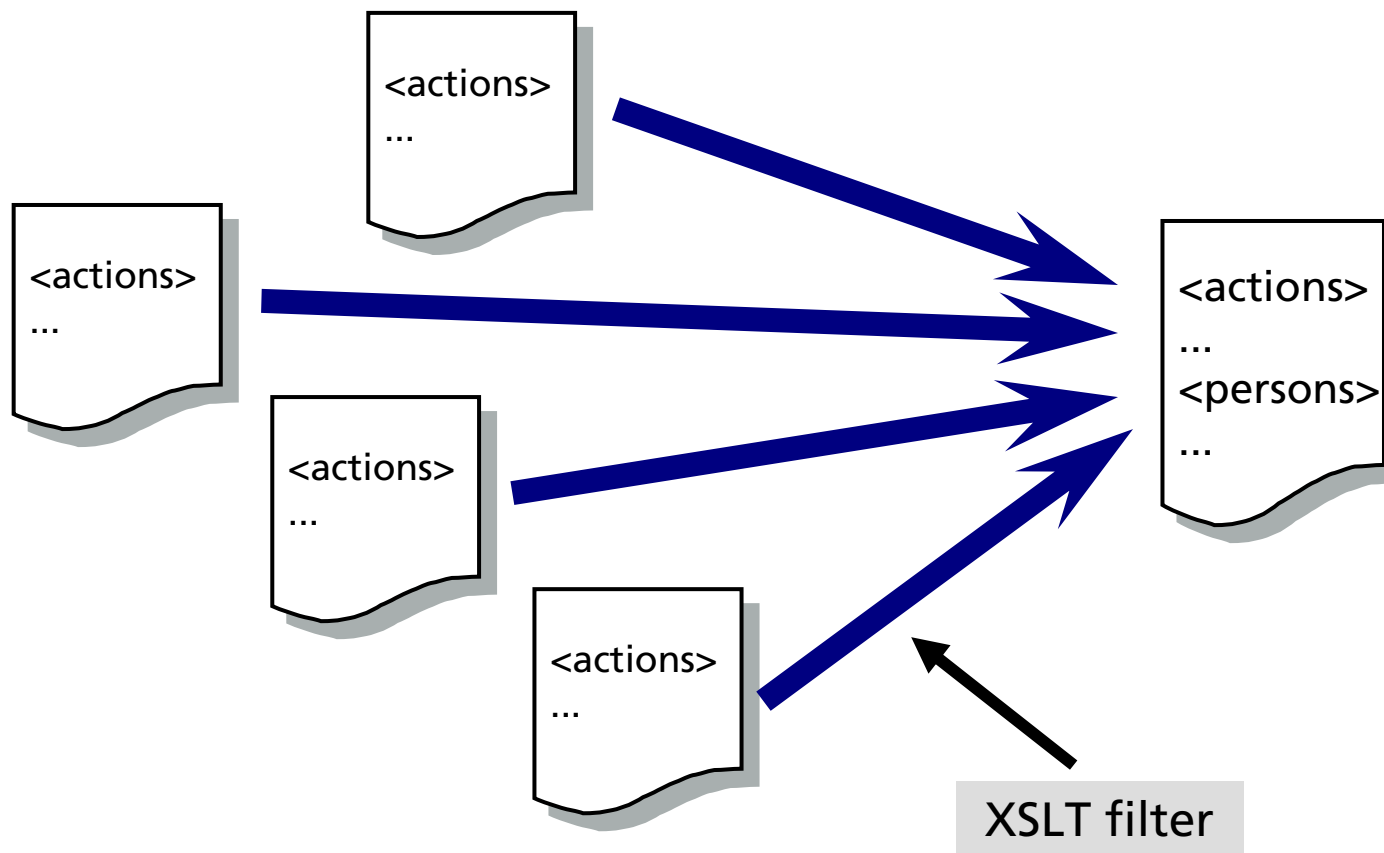
1.) The knowledge is automatically **extracted** and formally **represented**, ...



2. The Multilingual ZENON System - III

Dr. M. Hecking

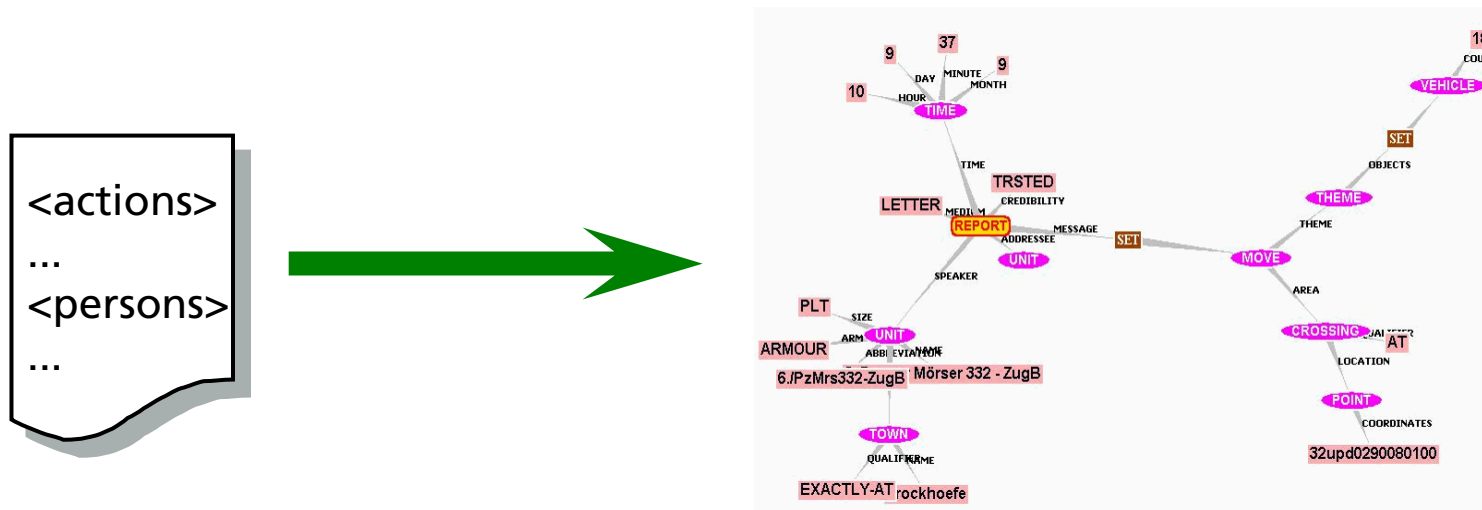
2.) ... then analysis-specifically **combined**, and...



2. The Multilingual ZENON System - IV

Dr. M. Hecking

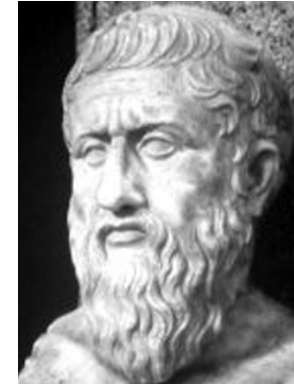
3.) ... and graphically presented.



2. The Multilingual ZENON System - V

Dr. M. Hecking

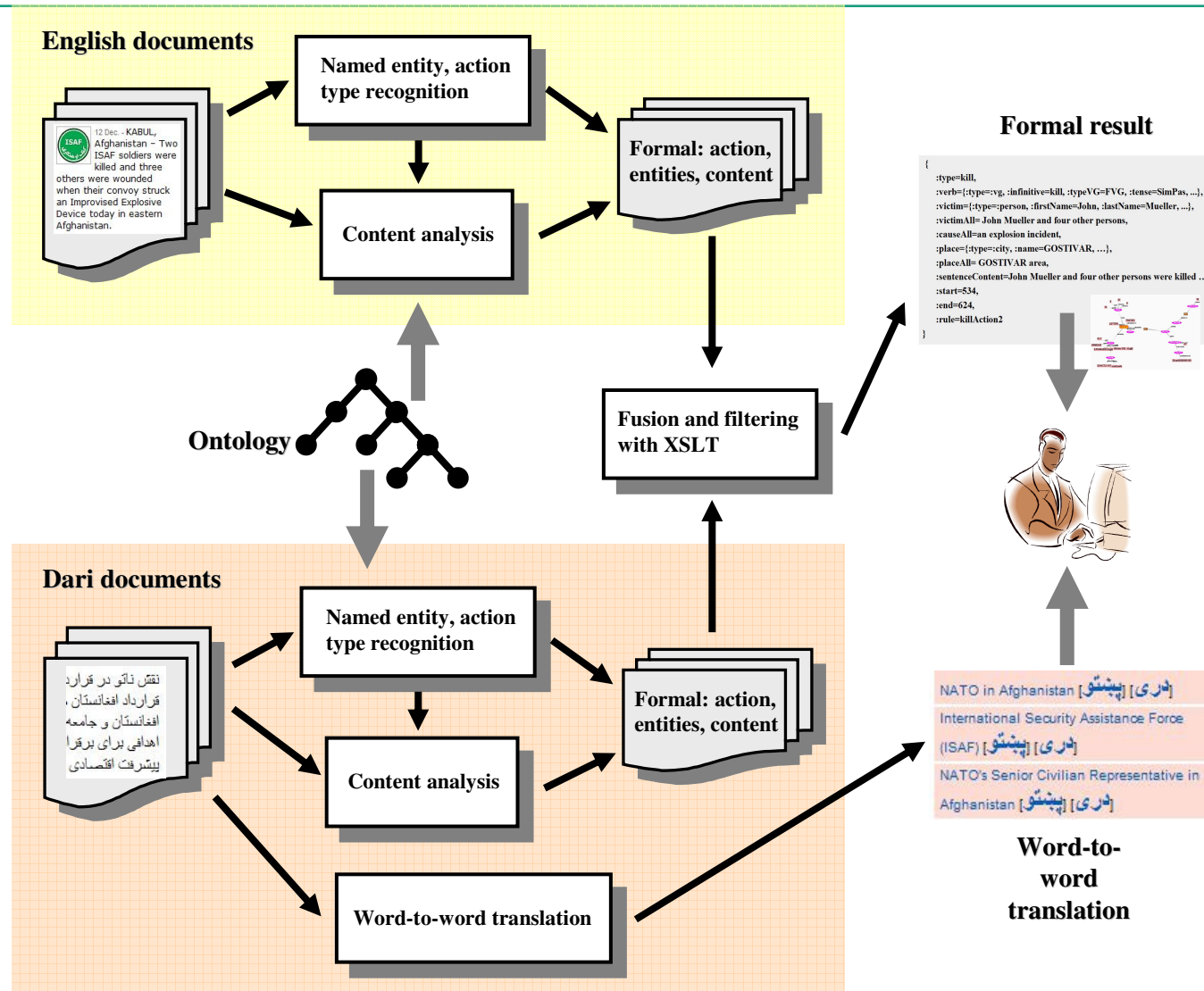
- Multilingual information extraction with the ZENON system, research prototype – not an operational system
- The information about the actions and named entities are identified from each sentence and the content of the sentences are formally represented in typed feature structures.
- These structures can be combined and presented in a graphically navigatable Entity-Action-Network.
- (Partial) information extraction from English HUMINT reports from the KFOR deployment, Dari texts, and Tajik texts.
- Also: a word-to-word-translation to further support the analyst.
- GATE: "is one of the most widely used human language processing systems in the world.", "comprises an architecture, framework (or SDK) and graphical development environment ...", University of Sheffield since 1995



Zenon of Citium
336 BC - 264 BC

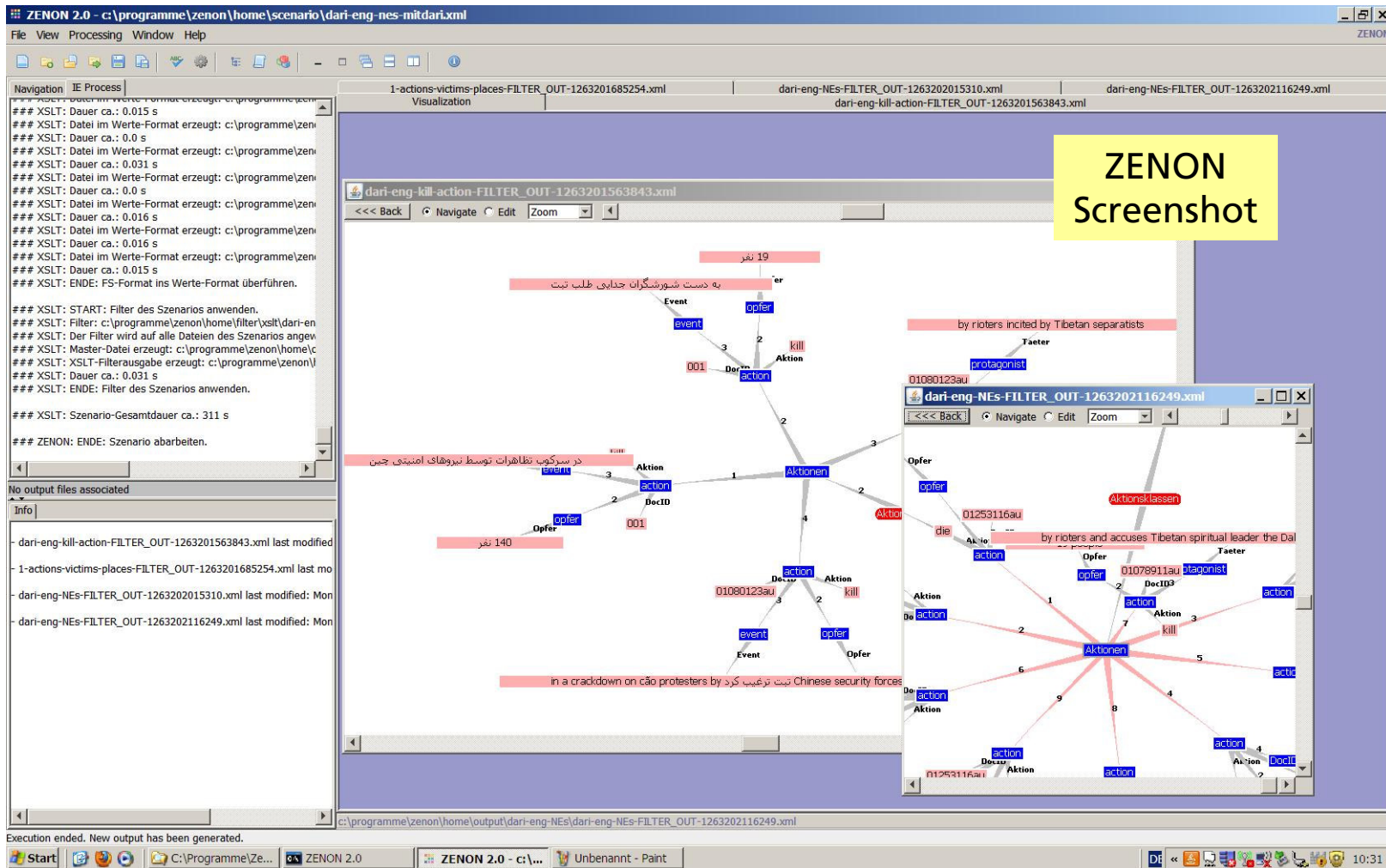
2. The Multilingual ZENON System - VI

Dr. M. Hecking



2. The Multilingual ZENON System - VII

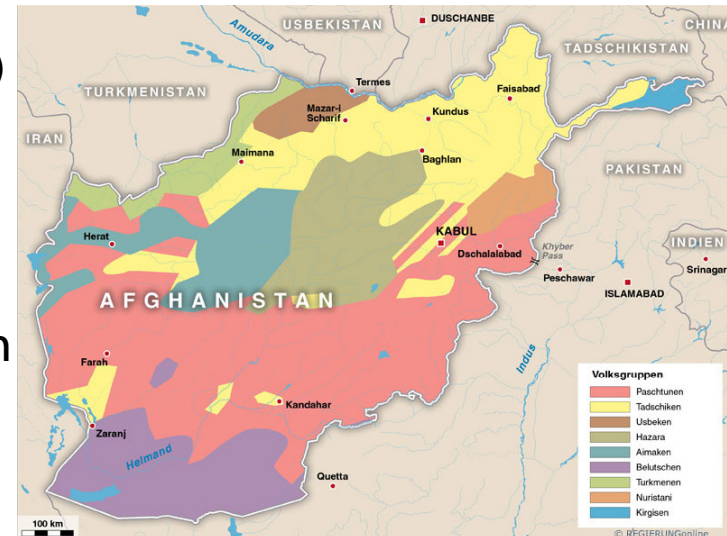
Dr. M. Hecking



3. The Multilingual Tajik Extension - I

Dr. M. Hecking

- The **Tajik** language (Tajik Persian, тоҷикӣ) is a modern version of Persian spoken in Central Asia.
- Most speakers: in Tajikistan, Uzbekistan
- Tajik is the official language of Tajikistan
- 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.



Женрол Азизаҳмади Вардак, 1
пулиси вилояти Пактиё ба Би
масъули нерӯҳои амниятии ви
атрофиёни ӯ дар ин инфиҷор
Пактиё шуморе аз афроди пул
кушта шудаанд.

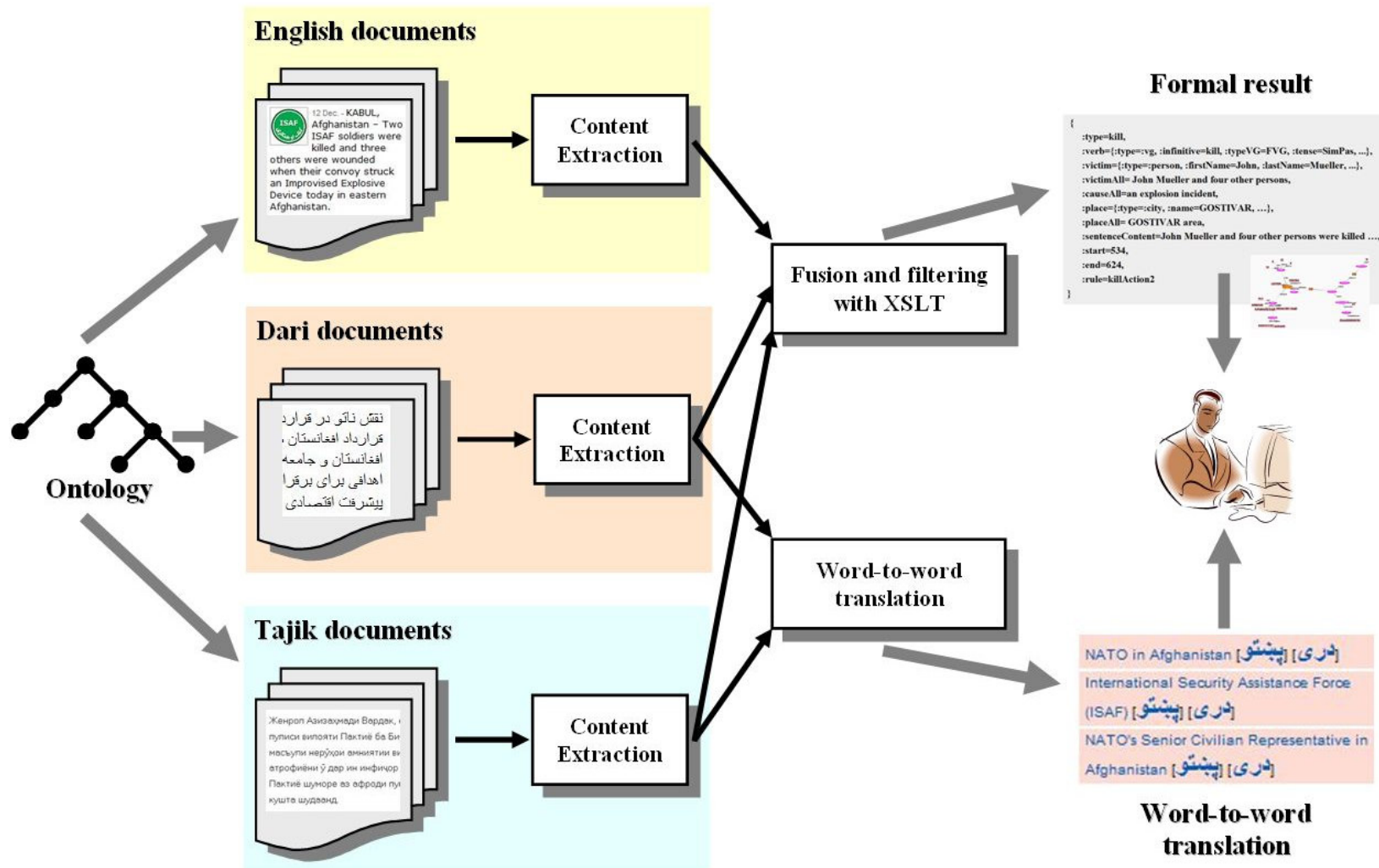
3. The Multilingual Tajik Extension - II

Dr. M. Hecking

- The ZENON research system will be extended by natural language processing functionality for *Tajik* texts.
- With the Tajik extension of the ZENON system it is possible to fuse information from sources written in the three different languages.
- The GATE module was already implemented, but it is not yet integrated into the overall system.
- Functionalities:
 - formal descriptions of named entities
 - formal descriptions of verbs
 - a word-to-word translation for the analyst (like the Dari module).
- Note: no combination of NEs and verbs at the moment (no assignment of semantic roles)

3. The Multilingual Tajik Extension - III

Dr. M. Hecking



3. The Multilingual Tajik Extension - IV

Dr. M. Hecking

- How the Tajik module was developed:
 - Documents were available in Tajik with their English and Russian translations
 - A broad collection of linguistically annotated Tajik texts (corpora) is not currently available.
 - Statistically-based machine learning methods are therefore not usable
 - The classical rule-based approach is used.
- Components of the Tajik module:
 - Tajik Tokenizer, Tajik Sentence Splitter, Gazetteer lists
 - Tajik NE Transducers
 - Tajik VP-Chunker
 - word-to-word translation into German

3. The Multilingual Tajik Extension - V

Dr. M. Hecking

- 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.
- All Gazetteer lists together contain about 2,100 entries (also some Gazetteer lists with English names)
- Gazetteer types with the number of entries:

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)	

3. The Multilingual Tajik Extension - VI

Dr. M. Hecking

- Based on the recognition of the Gazetteer list entries the **named entities** (NE) are identified with the help of the *Tajik NE Transducers*.
- These JAPE grammars generate **new annotations** that are used in subsequent processing steps, e.g. PERSON{firstname=Hamrokhon}
- Implemented NE **types**: *City, CommonOrg, Country, Date, GovernmentOrg, MilitaryOrg, Money, Person, Number, Gender, Jobtitle, Province, Region, TerroristOrg*.
- All types have various **features** which are also determined during the recognition process (e.g., *firstname, lastname, rule*).

3. The Multilingual Tajik Extension - VII

Dr. M. Hecking

The screenshot displays the GATE 4.0 build 2752 interface. The main text area shows a paragraph of Tajik text with several entities highlighted in red. A central dialog box titled 'PERSON' is open, showing a table with columns for 'Type', 'Set', 'Start', 'End', and 'Features'. The table contains three rows of entity data. The right sidebar shows a list of entity types with checkboxes, including 'DATE', 'PERSON', 'VERB', and 'VG'.

Type	Set	Start	End	Features
PERSON		22	36	{name=Мулло Абдулло, rule1=GazPersonFull, rule2=PersonFinal2}
DATE		39	49	{name=01.06.2009, rule1=TempYear, rule2=DateOnlyFinal}
PERSON		68	74	{firstname=Нагис, lastname=Хамробоева, rule=PersonFinal, rule1=FirstName, rule2=LastN}

- Named entity recognition of type **Person** and **Date**

3. The Multilingual Tajik Extension - VIII

Dr. M. Hecking

- The verbal phrases must be analyzed as a basis for the identification of the **action**.
- 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 (**partial** morphological analysis).
- Four **full-form lexica** for Tajik verbs
 - present-participle forms
 - past-participle forms
 - past-participle forms for compound verb phrases
 - compound verb phrases forms in past tense

3. The Multilingual Tajik Extension - IX

Dr. M. Hecking

- To identify the number, infinitive and the verb stem the approach of **word stemming** was used.
- The result of the analysis of simple verbs is stored in the annotation type **Verb**. Different features contain the recognized information:
 - infinitive ("Infinitiv"),
 - mood ("Merkmal")
 - person ("Person"),
 - stem ("Stamm"),
 - translation of the infinitive ("TranslationDE"),
 - tense ("Zeit"),
 - rule,
 - string.

3. The Multilingual Tajik Extension - X

Dr. M. Hecking

The screenshot displays the GATE 4.0 build 2752 interface. The left sidebar shows a project tree with 'Applications' containing 'Corpus Pipeline_00016', 'Language Resources' with 'GATE document_00040', and 'Processing Resources' including 'Tajik VP Chunker_00022', 'Tajik Transducer_00028', 'Tajik Sentence Splitter_0', 'Tajik Gazetteer_0001D', and 'Tajik Tokeniser_0001A'. The main window shows a text document with Tajik text. A 'Text' annotation window is open, displaying a table of verb annotations for the word 'шудан' (shudan).

Start	End	Features
867	875	{Infinitiv=шудан, Merkmal=Indikativ, Person=3.Pers.PI., ...}
937	944	{Infinitiv=шудан, Merkmal=Indikativ, Person=3.Pers.Sg., Stamm=шуд, TranslationDE=werden, Zeit=Perfekt, rule=W...
949	955	{Infinitiv=шудан, Merkmal=Indikativ, Person=3.Pers.Sg., Stamm=шуд, TranslationDE=sprechen,sagen, Zeit=Präsens, ...}

Verb annotations

3. The Multilingual Tajik Extension - XI

Dr. M. Hecking

The screenshot displays the GATE 4.0 build 2752 interface. The left sidebar shows a tree view with categories: Applications, Language Resources, Processing Resources, and Data stores. Under Language Resources, 'Corpus Pipeline_00016' and 'GATE document_0001A' are listed. Under Processing Resources, several Tajik-specific components are shown, including 'Tajik VP Chunker_00019'. The main window shows a text document with Tajik text. A pop-up window for 'VG' (Verb Group) annotations is visible, showing a table with columns: Type, Set, Start, End, and Features. The table lists several VG annotations with their corresponding features. A right-hand panel shows a list of annotation types with checkboxes, including DATE, DEFAULT_TOKEN, LOCATION, Lookup, LookupExtension, NUMBER, ORGANIZATION, PERSON, Sentence, SpaceToken, Split, Token, VERB, and VG (which is checked).

Type	Set	Start	End	Features
VG		860	876	{Merkmal=Passiv, Typ=FVG, Zeit=Präsens, rule=PresPassiv}
VG		485	498	{Merkmal=Passiv, Typ=FVG, Zeit=Präsens, rule=PresPassiv}
VG		226	229	{Infinitiv=хастан, Merkmal=Indikativ, Person=3.Pers.Sg, TranslationDE=sein, Typ=UVF, Zeit=Präsen}
VG		814	821	{Infinitiv=будан, Merkmal=Indikativ, TranslationDE=sein, Typ=UVF, Zeit=Präsens, rule=SeinPres}
VG		730	745	{Merkmal=Passiv, Typ=FVG, Zeit=Perfekt}

- Annotations of compound verb phrases (annotation type VG).

3. The Multilingual Tajik Extension - XII

Dr. M. Hecking

- Also a translation submodule to give a rough 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 this 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").

3. The Multilingual Tajik Extension - XIII

Dr. M. Hecking

Арабистони Саъудӣ:POS=NNP:TgWortanzahl=mw2:TranslationDE=Saudi Arabien
Артиллерист:POS=NN:TgWortanzahl=sw:TranslationDE=Artillerist
Аскарӣ 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
Афғонистон: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
Балчувон:POS=NNP:TgWortanzahl=sw:TranslationDE=Baldschuvon
Баҳри:POS=NN:TgWortanzahl=sw:TranslationDE=Meer
Баҳрнавард:POS=NN:TgWortanzahl=sw:TranslationDE=Matrose

■ Dictionary entries

3. The Multilingual Tajik Extension - XIV

Dr. M. Hecking

The screenshot shows the GATE 4.0 build 2752 interface. The left sidebar contains a tree view with categories: GATE, Applications, Corpus Pipeline_00016, Language Resources, Corpus for GATE documents, GATE document_00017, Processing Resources, Tajik VP Chunker_00022, Tajik Transducer_00021, Tajik Sentence Splitter_0, Tajik Gazetteer_0001D, Tajik Tokeniser_0001A, Document Reset PR_000, and Data stores.

The main window displays a text document with Tajik text. The text is annotated with various tags, including "Lookup". A "Lookup" table is visible, showing word-to-word translations. The table has columns for Type, Set, Start, End, and Features. The features column contains a list of features for each annotation, such as {POS=NNP, TgWortanzahl=mw4, TranslationDE=Nachrichtenagentur Asia Plus, majorType=word, minorType=tajik}.

Type	Set	Start	End	Features
Lookup		382	396	{POS=NNP, TgWortanzahl=mw4, TranslationDE=Nachrichtenagentur Asia Plus, majorType=word, minorType=tajik}
Lookup		382	396	{majorType=organization, minorType=tajik}
Lookup		386	395	{majorType=organization, minorType=tajik}
Lookup		415	420	{POS=NN, TgWortanzahl=sw, TranslationDE=Recht, majorType=word, minorType=tajik}

- The dictionary is implemented as a Gazetteer list.
- **Lookup annotations** with word-to-word translations.

4. Conclusion

Dr. M. Hecking

- In this presentation, we presented the functionality to perform information extraction for Tajik texts in the multilingual 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.
- Possible **improvements**
 - greater coverage of grammatical phenomena of Tajik
 - realize the recognition of action types and the combination of actions with their semantic roles
 - larger dictionary, independent translation system, deeper integration
- 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. References

Dr. M. Hecking

- M. Hecking. *Multilinguale Textinhaltserschließung auf militärischen Texten*. In: Verteilte Führungsinformationssysteme. Michael Wunder, Jürgen Grosche (Hrsg.), Springer-Verlag, 2009.
- M. Hecking, C. Schwerdt. *Multilingual Information Extraction for Intelligence Purposes*. In: Proceedings of the 13th International Command and Control Research and Technology Symposium (ICCRTS), June 17-19, 2008, Bellevue, WA, U.S.A.
- M. Hecking. *System ZENON – Semantic Analysis of Intelligence Reports*. In: Proceedings of the LangTech 2008, February 28-29, 2008, Rome, Italy.
- C. Schwerdt. *Analyse ausgewählter Verbalgruppen der Sprache Dari zur multilingualen Erweiterung des ZENON-Systems*. FGAN, FKIE-Bericht Nr. 146, 2007.
- T. Sarmina-Baneviciene. *Analyse spezifischer Probleme der tadschikischen Sprache zur multilingualen Erweiterung des ZENON-Systems*. Fraunhofer FKIE, 2010 (forthcoming).

Thank you for your attention!



Questions?