

---

# Fast Realization of Automatic Translation Systems for New Mission-Relevant Languages

---

**Dr. Matthias Hecking**  
**Sandra Noubours**

Fraunhofer Institute for Communication,  
Information Processing and Ergonomics FKIE  
Neuenahrer Straße 20, 53343 Wachtberg, Germany  
matthias.hecking@fkie.fraunhofer.de  
sandra.noubours@fkie.fraunhofer.de

1. Introduction
2. Rough translation
3. Concept for the rapid realization of translation systems
4. Language pair Dari – German as an example
5. Conclusion, References

# 1. Introduction – I

Berisha: Datën e zgjedhjeve nuk ndry...

Kryeministri Sali Be... deklaroi sot në mb... Grupit Parlamentar... zgjedhjet e datës... do të ndryshojnë p... qendrimeve të kre... opozitës.

Ai theksoi se opoz... jepen të gjitha garancitë për zhvillimin normal të zg...

"Çdo ditë zoti Rama kupolës së tij i thotë se në nuk do të marrim pjesë në zgjedhje dhe në do zvarnisim vendin si në 1997-n. Zgjedhjet do të bëhen në 8 maj dhe nuk do ketë forcë që do i zhvendosë ato. Opozita do ketë të gjitha hapësirat që duhet të ketë në një garë elektorale të një vendi të lire", tha kreu i qeverisë.

یوش متقال وزیر دفاعه یلوح للمالکي بمحدودية الصبر

فومانة ولایت هدف جنین

میثد. از اینرو، نیروهای اردو و پولیس تصمیم این بار اقدام مذکور یک عمل متفاوت به نظر می آید. المللی به عهده نداشتند. این افسران افغانی بود از آغاز الی بیاده نمودن عملیات را در ساحه به عهده داشتند. همه اولتر پلانگذار ی نمودند.

واخلی. تونر جیمی آن وویل: "اوس نوم او ددی بنوونخی د چارو د یرمخ و مشوری اولار بنوونی. هغه هم که د نومرو مراسم وروسته، په ژوند تر لاسه شوی وه، ننداری ته وړاندی

道原

老子曰：「有物混成，先天地，不聞其聲，吾強為之名，字可測，苞裹天地，莫受無形，清，施之無窮，無所朝夕，表柔而能剛，含陰吐陽，而章三走，鳥以之飛，麟以之遊，鳳車取尊，以退取先。古者三以撫四方。是故能天運地，始。風興雲蒸，當聲雨降，並

همه اولتر پلانگذار ی نمودند.

واخلی.

تونر جیمی آن وویل: "اوس نوم او ددی بنوونخی د چارو د یرمخ و مشوری اولار بنوونی. هغه هم که د نومرو مراسم وروسته، په ژوند تر لاسه شوی وه، ننداری ته وړاندی

همه اولتر پلانگذار ی نمودند.

واخلی.

تونر جیمی آن وویل: "اوس نوم او ددی بنوونخی د چارو د یرمخ و مشوری اولار بنوونی. هغه هم که د نومرو مراسم وروسته، په ژوند تر لاسه شوی وه، ننداری ته وړاندی

- Military operations: documents written in foreign languages are relevant.
- The information in the documents might be of great value for the military analyst.
- Foreign languages can be an obstacle.

- Problem if the documents are written in **less-learned languages**:
  - only a few or no human translators are available, security issues
  - no economic interest in building translation systems
- New deployments or new languages for intelligence purposes: **How** the military system can **react agile to this language problem?**
- This paper: Propose a **concept** to improve this situation:
  1. **Reducing** the expectations of the **quality** of the translation and
  2. using the approach of **statistical machine translation (SMT)** to rapidly produce new translation system

- Gisting: Possible to rapidly construct SMT systems if **rough translations** are sufficient .
- Gisting:
  - to understand the general content or
  - identify those documents which should be translated by human translators.
- As an example, we describe how we used this approach to set up an SMT system for the language pair **Dari – German**.

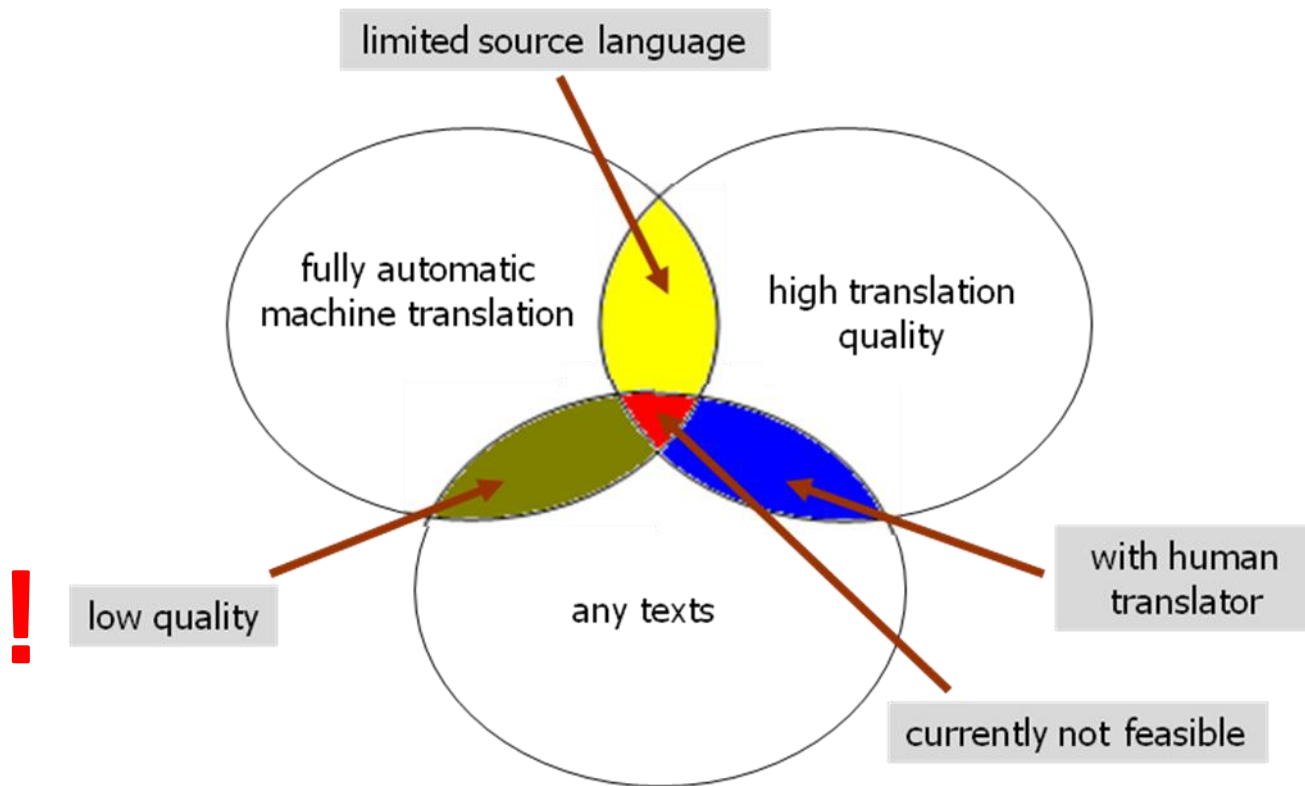
## 2. Rough translation – I

---

- **Machine translation (MT)** is the complete automatic translation of text from one natural (source) language to another (target language) while preserving the meaning.
- Not: Computer-aided translation used by humans (translation memories).
- **Gisting:** rough translation, not a high quality translation, possible wrong translated words, grammar errors etc.

## 2. Rough translation – II

### ■ Different types of translation



The graphic is from the talk "Machine Translation II" given by Harold Somers, School of CS, University of Manchester.

## 2. Rough translation – III

---

- For **new** mission-relevant languages **any text** written in this language might be of interest. Therefore, we have to accept the **low quality** of the translation.
- According to previous slide this is doable by **fully automatic** machine translation systems.
- But, if we want to **adapt agile** to new “language-situations” this is only possible if the systems for fully automatic translations can be constructed **rapidly**.



## 2. Rough translation – IV

- Example for rough translation systems:
  - Forward Area Language Converter (FALCon, U.S. Army Research Laboratory)
    - notebook-based
    - documents are scanned, OCR (optical character recognition)
    - translated into English
    - English text can be searched for keywords
    - used during the Haiti (1995) and the Bosnia (1997) mission



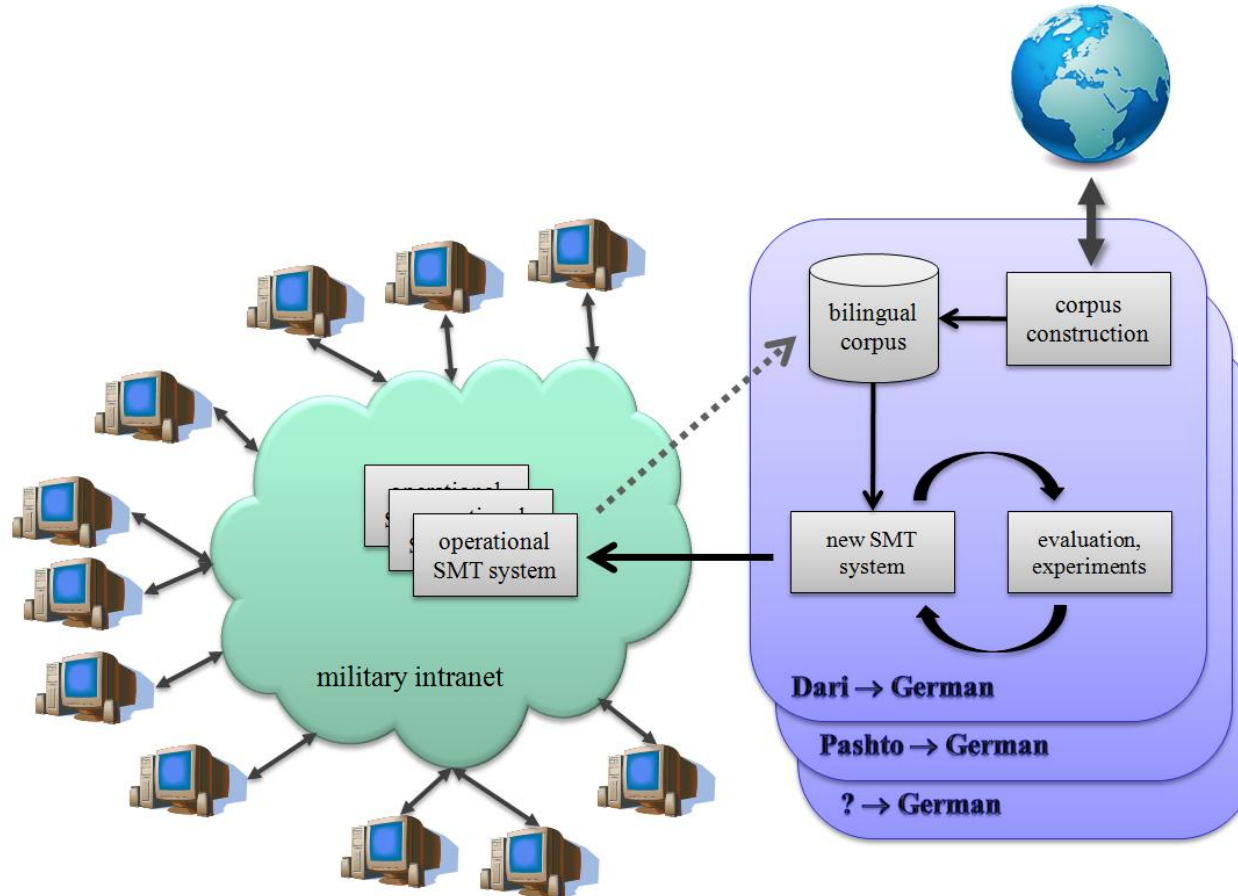
### 3. Concept for the realization of translation systems – I

---

#### ■ Parts of the (simple) concept:

1. Reduce the expectations concerning the quality of the automatic translation to **rough translation**.
2. Use the approach of statistical machine translation (**SMT**) to come up very fast with a new translation system.
3. Build up a **team of scientists** who are up-to-date to the SMT technology and to the corresponding scientific field and who are responsible to create very fast new versions of the translation system.
4. Create an military operational centralized automatic **translation service** for the military users via any military intranet.
5. Make sure that the operational staff **tightly works together** with the team of scientists.

# 3. Concept for the realization of translation systems – II



## 4. Dari – German as an example - I

---

- SMT is crucial for the success of the concept.
- We build up a **small infrastructure** (small team of scientists, computer cluster, procedures, software ...) to show that we can come up (fast) with a new SMT system for a language pair relevant for the Bundeswehr.
- Project: **Machine Translation for ISAF Forces (ISAF-MT)**:
  - Objectives: Build up **bilingual corpora** and **linguistic tools** and to construct through SMT technology **Dari – German translation systems**
  - German – U.S. cooperation project (Air Force Research Laboratory, Dari – English)
- Our project proved that a translation system can be produced rapidly (depending of the availability of corpora).

## 4. Dari – German as an example - II

---

### ■ Dari

- spoken by approx. 22 million Afghans
- spoken in the center and in the northern part of Afghanistan
- an Indo-European language
- 28 Arabic characters + 4 additional
- a right-to-left language, no distinction between uppercase and lowercase, no written short vowels
- syntactic word order SOV (subject-object-verb)
- closely related to the modern Iranian Persian (Farsi) and Tajik spoken in Tajikistan

## 4. Dari – German as an example - III

### ■ Statistical machine translation (SMT)

- goal: find the “most likely” (best) translation for a given source language sentence
- for the source language sentence  $f$  the target language sentence  $e$  is selected that maximizes the probability  $p(e|f)$

$$\arg \max_e p(e|f) = \arg \max_e \frac{p(e)p(f|e)}{p(f)}$$

- **translation model**  $p(f|e)$ : retains the content of the source language sentence, bilingual parallel corpora
- **language model**  $p(e)$ : the target language sentence is well-formed, monolingual corpus

## 4. Dari – German as an example - IV

### ■ Excerpt from the ISAF-MT translation model

```
154229 ، ساکت باتنید ، ||| ، sei still ! ||| 0.197548 0.00202913 0.197548 0.00328463 ||| ||| 1 1
154230 ، ساکت باتنید ، ||| ، sei still ||| 0.197548 0.00331125 0.197548 0.00420655 ||| ||| 1 1
154231 ، ساکت Welt . online می نویسد : درمطلبی با عنوان می نویسد : ||| Quelle Welt online am 20. Oktober 2008 : ||| 0.0282211
154232 ، ساکت Welt . online می نویسد : درمطلبی با عنوان می نویسد : ||| Quelle Welt online am 20. Oktober 2008 ||| 0.0282211 4.
154233 ، ساکت های ، ||| Stationierung von ||| 0.00581022 3.81773e-06 0.0246934 7.41779e-06 ||| ||| 34 8
154234 ، ساکت های ، ||| Stationierung ||| 0.00548743 3.81773e-06 0.0246934 0.0006259 ||| ||| 36 8
154235 ، ساکت های ، ||| begleitet von der Stationierung von ||| 0.00731657 3.81773e-06 0.0246934 2.60816e-13
154236 ، ساکت های ، ||| begleitet von der Stationierung ||| 0.00731657 3.81773e-06 0.0246934 2.20072e-11 ||| |||
154237 ، ساکت های ، ||| der Stationierung von ||| 0.00731657 3.81773e-06 0.0246934 4.73274e-07 ||| ||| 27 8
154238 ، ساکت های ، ||| der Stationierung ||| 0.00731657 3.81773e-06 0.0246934 3.9934e-05 ||| ||| 27 8
```

### ■ Excerpt from the ISAF-MT language model

```
-0.9569345 für die Dauer des -0.140566
-0.3309343 für die Dauer von 0.0009403285
-1.329289 nur die Dauer von -0.006186663
-0.4907208 oder die Dauer der -0.006186663
-0.4907208 und die Dauer der -0.006186665
-0.4907208 war die Dauer der -0.2274503
```

## 4. Dari – German as an example - V

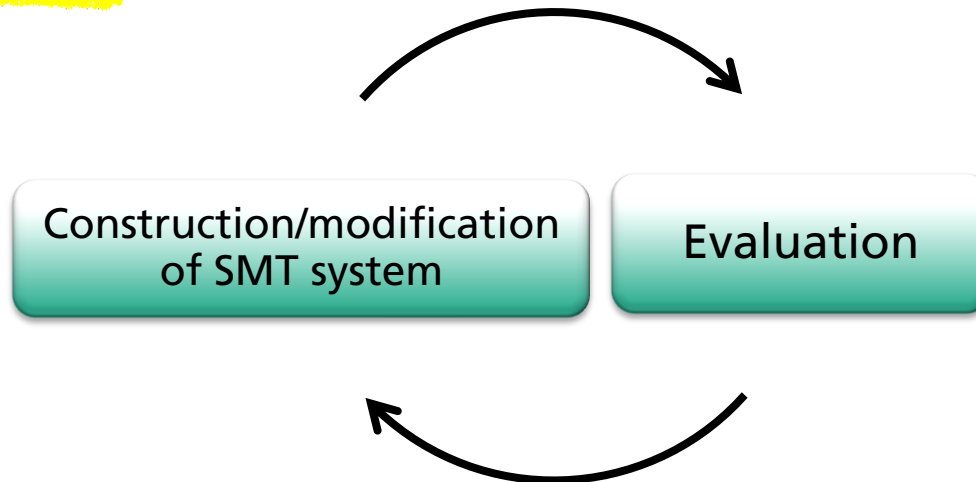
- The translation model is trained on a **bilingual parallel corpus**
- Dari – German corpus: topics “military” and “terrorism”; Sada-e-Azadi, Pajhwok, Kokchapress; around 27,000 sentences; military dictionary Dari to German (71,798 entries)

|    |   |    |  |
|----|---|----|--|
| 1  | دوی نیروهای امنیتی افغان ثابت کردند که توانایی تامین فرقه جدید ۱۱۱ امنیت کابل را تقویت می بخشد بعد از اینکه | 1  | Die 111. Division unterstützt die Sicherheitskräfte in Kabul. Nachdem afgha  |
| 2  | وزارت دفاع ملی در اخیر ماه می فرقه جدید ۱۱۱ اردوی ملی افغان را در کابل تاسیس نمود.                          | 2  | Ende Mai wurde die 111. Division vom afghanischen Verteidigungsministe       |
| 3  | این فرقه با همکاری نزدیک با دیگر نیروهای ملی امنیت افغان چون یولیس و امنیت کار خواهد نمود.                  | 3  | Die Division wird in Kooperation mit anderen afghanischen Sicherheitskrä     |
| 4  | یخت می باشد، فرقه ۱۱۱ مسئولیت امنیت شهر را زمانی بدوش خواهد گرفت که کابل را خطر بزرگی تهدید نماید           | 4  | Während die afghanische Nationalgarde hauptverantwortlich für die Vertei     |
| 5  | دم خدمت خواهیم نمود   | 5  | "Ich verspreche der afghanischen Regierung, dass wir unserem Volk unte       |
| 6  | ء و پنج کندک می باشد  | 6  | Kabuls 111. Division, die ihr Training vor sechs Monate begonnen hat, bes    |
| 7  | کنندگان مجهز می شوند  | 7  | Die Soldaten besuchen ein zweimonatiges Training und ihre modernen Wi        |
| 8  | تیز قواء مسلح وزارت   | 8  | Die Gründung der neuen Division war ein großer Schritt in der weiteren Ei    |
| 9  | زی را فراهم نموده اند   | 9  | "Die afghanische Armee hat mehr Sicherheit in die afghanischen Provinze      |
| 10 | ی خارجی کار می کنند   | 10 | Weiterhin kündigte er eine neue Militärinitiative an, die aus Polizeibeamter |
| 11 | مه بین المللی شده است   | 11 | Die letzten Vorfälle, in denen Zivilisten aus Versehen ums Leben gekomme     |
| 12 | قدامت می باشد بهبود   | 12 | Die sieben Jahre alte afghanische Armee besteht mittlerweile aus 90.000 §    |



## 4. Dari – German as an example - VI

- Development of SMT systems: **experimental approach**, different ways of realizing the translation or language model, different corpus pre-processing, ...
- Evaluation is based on a manually checked bilingual corpus (around 1,000 phrases); various metrics, e.g., BLEU, METEOR, TER
- Toolbox: **Moses**



## 4. Dari – German as an example - VII

---

- **Overall objective of the experiments:** Find improvements of the translation model (and its submodels) and correct weights of the parameter in the models that maximizes the probability of produced translated sentences.
- Considered parameters:
  - size and quality of the tuning set,
  - normalization, compound splitting,
  - reversal of sign and word order,
  - alignment heuristics,
  - maximum phrase length,
  - reordering,
  - inclusion of part-of-speech (POS),
  - inclusion of lemma,
  - inclusion of word stems.

## 4. Dari – German as an example - VIII

- Experiment: **Compound splitting**
- The German language has a lot of compound words; words consisting of more than one stem, e.g., “darkroom”
- In this experiment the compounds were split by the Moses Compound Splitter into single words; for both languages

| Compound splitting              | BLEU | METEOR | TER   |
|---------------------------------|------|--------|-------|
| baseline                        | 7,55 | 15,60  | 8,93  |
| input (dari)                    | 7,14 | 15,54  | 9,02  |
| output (deutsch)                | 6,65 | 14,98  | 10,50 |
| input (dari) & output (deutsch) | 7,14 | 15,60  | 9,02  |

- **Decline** in the scores.
- Contrary to the expectation compound splitting does not increase the quality of the translation.

## 4. Dari – German as an example - IX

- Experiment: **Reversal of sign and word order**
- Dari has a right-to-left writing system and the verb is located at the end of the sentence.
- In this experiments each line of the Dari side was reversed sign by sign and word by word.

| Reversal       | BLEU  | METEOR | TER  |
|----------------|-------|--------|------|
| baseline       | 11,10 | 16,55  | 8,87 |
| reverse words  | 10,03 | 15,48  | 9,41 |
| reverse string | 10,07 | 15,50  | 9,38 |

- **Decline** in the scores.
- Simple reversing is not an appropriate way of handling the direction in different writing systems.

## 4. Dari – German as an example - X

- Experiment: **Inclusion of lemma**
- Lemmas are abstractions of the various possible word forms, e.g. runs => run, läuft => laufen; integrated by means of **factored translation models**.
- German: Mate-Tools package; Dari: internal tool.

| Inclusion of lemma                            | BLEU        | METEOR       | TER         |
|---|-------------|--------------|-------------|
| <b>Baseline</b>                               | <b>5,34</b> | <b>12,14</b> | <b>9,11</b> |
| <b>align: lemma =&gt; lemma</b>               | 5,77        | 12,60        | 9,03        |
| <b>align: word =&gt; lemma</b>                | 5,60        | 12,14        | 9,02        |
| <b>align: word + lemma =&gt; lemma</b>        | 5,61        | 12,23        | 9,02        |
| <b>trans: lemma =&gt; word + lemma</b>        | 5,37        | 12,18        | 9,06        |
| <b>trans: word =&gt; word + lemma</b>         | 5,45        | 12,18        | 9,09        |
| <b>trans: word + lemma =&gt; word</b>         | 5,35        | 12,15        | 9,08        |
| <b>trans: word + lemma =&gt; word + lemma</b> | 5,36        | 12,13        | 9,09        |
| <b>reord: lemma =&gt; word</b>                | 5,40        | 12,32        | 9,07        |
| <b>reord: word + lemma =&gt; word</b>         | 5,42        | 12,26        | 9,09        |

- Alignment can **benefit** from the inclusion of lemma.

## 4. Dari – German as an example - XI

### ■ Experiment: Example outputs of the ISAF-MT translation system

**Input:** چهل یک برکت است . **What I do not know will not hurt me.**  
**Reference:** Was ich nicht weiß , macht mich nicht heiß .  
**Output:** Unkenntnis ein Segnung . **semantic affinity**

**Input :** بیش از 100 هزار سرباز امریکایی در افغانستان مستقر هستند که علیه طالبان و شبکه تروریستی القاعده می جنگند .  
**Reference :** Über 100.000 US - Soldaten sind in Afghanistan stationiert , die gegen die Taliban und das al - Qaida Terror - Netzwerk kämpfen .  
**Output :** mehr als 100.000 US - Soldaten Soldaten in Afghanistan stationiert , gegen Taliban und dem Al-Qaida Terror - Netzwerk kämpfen . **semantic similarity**

**Input :** من عجله دارم . **small deviation**  
**Reference :** Ich bin in Eile .  
**Output :** ich hab ' s eilig . **I'm in a hurry.**

**Input :** ان سی دی 10 دلار قیمت دارد . **identity**  
**Reference :** Diese CD kostet 10 Dollar .  
**Output :** Diese CD kostet 10 Dollar .  
**This CD costs \$ 10.**

- Understanding documents written in foreign languages are important in preparation of military operations or during these operations.
- Less-learned language are a problem.
- To overcome the unsatisfying situation we propose a concept:
  - Reduce the expectations for the quality of the translation.
  - Use SMT to rapidly produce new translation systems.
- As an example of using the SMT technology we show a translation system for the language pair Dari – German.
- We report about the corpus construction and the experiments to improve the translation system.
- We were successful in realizing an SMT system of a language pair relevant for the needs of the Bundeswehr. We were able to do this in a couple of month.

## 5. References

- M. Hecking, S. Noubours. *Machine Translation for ISAF Forces (ISAF-MT)*. Fraunhofer-Institut für Kommunikation, Informationsverarbeitung und Ergonomie FKIE, Wachtberg, Germany, Dokumentation zum Forschungsvorhaben E/IB1S/AA166/9F008, 31. Dezember 2011 (in German).
- M. Hecking, S. Noubours. *Machine Translation for ISAF Forces (ISAF-MT)*. Fraunhofer-Institut für Kommunikation, Informationsverarbeitung und Ergonomie FKIE, Wachtberg, Germany, Dokumentation zum Forschungsvorhaben E/IB1S/AA166/9F008, 31. Dezember 2010 (in German).
- M. Hecking, A. Wotzlaw, R. Coote. *Multilingual Content Extraction Extended with Background Knowledge for Military Intelligence*. In: Proceedings of the 16th International Command and Control Research and Technology Symposium (ICCRTS), June 21-23, 2011, Québec City, Québec, Canada.
- M. Hecking, T. Sarmina – Baneviciene. *A Tajik Extension of the Multilingual Information Extraction System ZENON*. In: Proceedings of the 15<sup>th</sup> International Command and Control Research and Technology Symposium (ICCRTS), June 22-24, 2010, Santa Monica, CA, U.S.A.
- 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?**