# ACCURAT

Analysis and Evaluation of Comparable Corpora
for Under Resourced Areas of Machine Translation

www.accurat-project.eu

**Project no. 248347**

# Deliverable D5.2
# Evaluation of Machine Translation for Narrow Domain

**Version No. 1.0**
**29/06/2012**

## Document Information

| | |
|---|---|
| Deliverable number: | D5.2 |
| Deliverable title: | Evaluation of Machine Translation for Narrow Domain |
| Due date of deliverable: | 30.06.2012 |
| Actual submission date of deliverable: | 29.06.2012 |
| Main Author(s): | Gr. Thurmair, V. Aleksić (LT), M. Tadic (FFZG) |
| Participants: | LT, FFZG, Tilde |
| Internal reviewer: | Tilde |
| Workpackage: | WP5 |
| Workpackage title: | Evaluation of usability in applications |
| Workpackage leader: | Tilde |
| Dissemination Level: | **PP**: Restricted to other programme participants (including the Commission Services) |
| Version: | V1.0 |
| Keywords: | machine translation, narrow domain, evaluation |

## History of Versions

| Version | Date | Status | Name of the Author (Partner) | Contributions | Description/ Approval Level |
|---|---|---|---|---|---|
| V0.1 | 01/06/2012 | Draft | LT | Initial draft | Delivered for review |
| V1.0 | 29.06.2012 | final | FFZG | FFZG part added | Submitted |

## EXECUTIVE SUMMARY

The task consists in the evaluation of improvements of MT systems (both RMT and SMT) if adapted to a narrow domain. Baseline systems were adapted to a narrow domain (automotive), and the improvements were measured by both automatic and human evaluation.

The results show that slight improvements can be achieved for German-English, while for English-Croatian a deterioration is measured. The reasons for this behaviour need to be explored further.

# Table of Contents

# Abbreviations

| Abbreviation | Term/definition |
|---|---|
| MT | Machine Translation |
| SMT | Statistical Machine Translation |
| TM | Translation Model |
| LM | Language Model |
| BLEU | Bilingual Evaluation Understudy |
| AC | Acquis Communautaire |
| EU | European Union |
| TU | Translation Unit |
| WMT | Workshop for Statistical Machine Translation |
| MERT | Minimal Error Rate Training |

# 1. Task / Objective

The objective of this subtask is to collect information on improvements which can be achieved by tuning Machine Translation systems for narrow domains, using data from comparable corpora. To do this, such data must be collected; these data must be used to improve a baseline MT system, and the adapted system must be compared with the baseline system.

In the present case, it was decided to use the automotive domain as an example for a narrow domain; and in order to limit this further, the subdomain on transmission / gearbox technology was selected for German-English pair, while general automotive domain and renewable energy was selected for English-Croatian pair. In order also to evaluate if different architectures of MT systems favour domain adaptation more or less, both a data driven (SMT) and a knowledge-driven (RMT) system were used for German-English pair, while only a data driven (SMT) system was used for English-Croatian pair.

Evaluation languages for the tests described here were:

1. German-to-English, where automotive technology play a significant practical role, and automotive players form a major part of (Linguatec) MT customers
2. English-to-Croatian with automotive technology and renewable energy narrow domains, since we expect that these narrow domains will exhibit a growth in translation in the following years.

# 2. Evaluation Objects: Narrow-domain-tuned MT systems

The object of the evaluation were two types of systems for German-English pair:

- **DFKI-adapted** is the automotive SMT system as created by DFKI (cf. Deliverable D4.3), based on standard SMT technology (GIZA++ and MOSES)
- **PT-adapted** is the automotive RMT system created from Linguatec's 'Personal Translator' PT (V.14), which is a standard rule-based MT system based on the IBM slot-filler grammar technology (Aleksić / Thurmair 2011)

These two evaluation objects were created as follows:

For the **baseline** systems, the 'Personal Translator' (V14) was taken as out of the box and installed on a standard PC. For 'SMT', DFKI trained the baseline with standard parallel data (Europarl, JRC etc.), as well as some initial comparable corpus data as collected in the first phase of ACCURAT (Deliverable D3.1).

The object of the evaluation for English-Croatian pair were two SMT systems adapted for two different narrow domains:

- automotive technology
- renewable energy

These objects were created by DFKI as they trained the baseline with standard parallel data (SETimes, Croatian-English Parallel Corpus etc.), as well as some initial comparable corpus data as collected in the first phase of ACCURAT (Deliverable D3.1).

## *2.1 Domain adaptation*

For the adapted versions of German-English pair, data were collected from the automotive domain. These data were made available by crawling sites of automotive companies being active in the transmission field (like ZF, BASF, Volkswagen and others). These data were strongly comparable. They were then aligned and cleaned manually. Some sentence pairs were set aside for testing, the rest was given to the two systems for domain adaptation as development and test sets. The resulting narrow-domain automotive corpus has about 42.000 sentences for German-to-English.

For the adapted versions of English-Croatian pair, data were collected from automotive domain and renewable energy domain. The Croatian automotive corpus was produced as the subset of hrWaC (Ljubešić / Erjavec 2011) comprising 5.7 Mw. For the English side, the English part of the automotive domain corpus collected for German-English pair was used. The renewable energy domain comparable English-Croatian corpus was crawled by ILSP with the focussed crawler.

For **statistical** MT, German-English domain adaptation was done in case of DFKI-adapted by adding the sentences to the training and development sets. A new version was created (called 'DFKI-adapted' and submitted to Linguatec for test in Q4/2011. In the case of the English-Croatian pair, both comparable corpora were processed with LEXACC tool and extracted parallel data were used by DFKI for two different domain adaptations of the existing English-Croatian baseline model.

In case of **rule-based** technology, domain adaptation is more complicated, as it involves terminology creation as main source of adaptation. Therefore, the following steps were taken:

- creation of a phrase table with GIZA++ and MOSES containing the translation of relevant terms and phrases; for this, the DFKI-adapted phrase-tables as well as phrase-tables created only from automotive sentences were used
- extraction of bilingual terminology candidates from these phrase tables using the P2G (Phrase-Table-to-Glossary) tool developed for this purpose; this resulted in a candidate list of about 25.000 term candidates
- preparation of these candidates for dictionary import, including creation of part-of-speech and gender annotations, removal of already existing entries, removal of candidates which could not be imported, resolution of conflicts in transfers etc.; the final list of imported entries was about 7100 entries.
- these entries were collected in a special user dictionary, which can be added to the system dictionary in cases where certain narrow domains need to be translated, using a special 'automotive' domain tag.

This procedure is described in detail in the ACCURAT Deliverable 4.4, as well as in (Thurmair/Aleksić 2012).

Result of these efforts for German-English pair were two test systems, called '*DFKI-adapted*' and '*PT-adapted*', both for German-to-English, and both tuned for automotive domain with the same adaptation data; they form the test objects of the present evaluation.

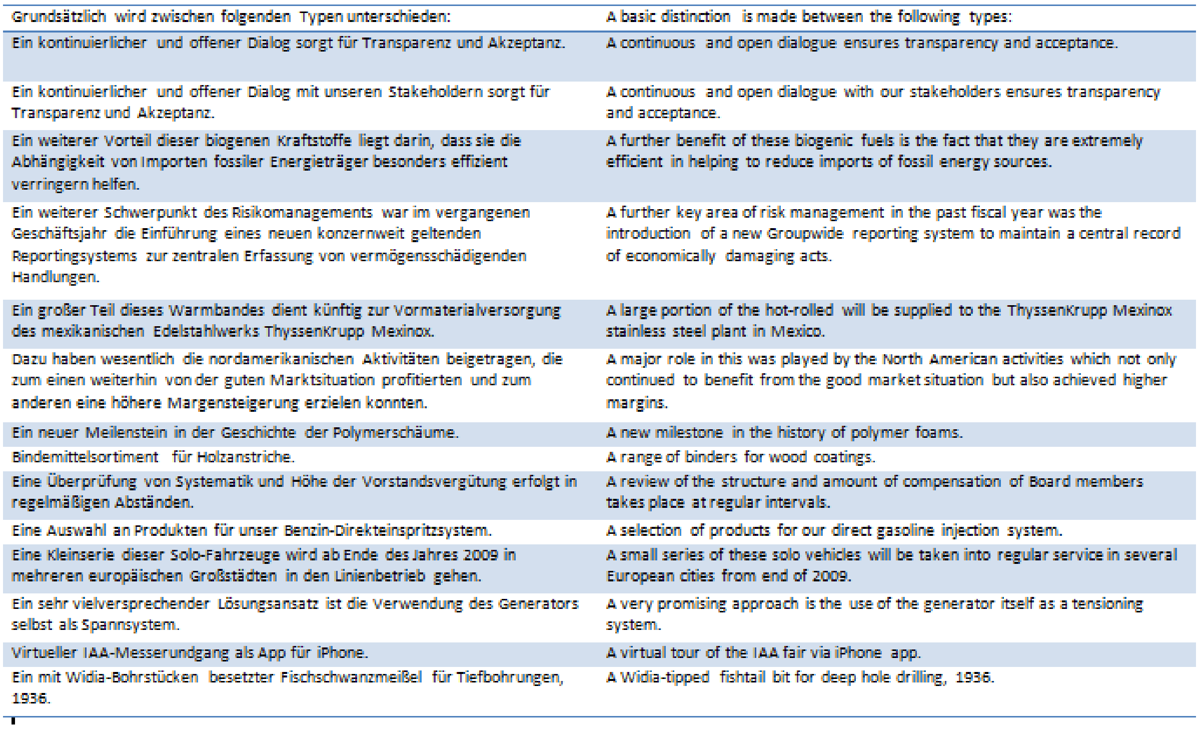
In the case of English-Croatian pair, two test systems were called '*enhr-automotive-adapted*' and '*enhr-REn-adapted*'.

# 3. Evaluation Data

For evaluation of German-English pair, a set of sentence pairs was extracted from the collected strongly comparable automotive corpora. In total about 1500 sentences were taken for tests, with one reference translation each.

The sentences represent 'real-life' data; they were not cleaned or corrected, just like the training data. So they contain spelling mistakes, segmentation errors and other types of 'noise'. This fact of course affects the translation quality for the adapted systems.

Examples are given in Figure 1.



| Grundsätzlich wird zwischen folgenden Typen unterschieden: | A basic distinction is made between the following types: |
|---|---|
| Ein kontinuierlicher und offener Dialog sorgt für Transparenz und Akzeptanz. | A continuous and open dialogue ensures transparency and acceptance. |
| Ein kontinuierlicher und offener Dialog mit unseren Stakeholdern sorgt für Transparenz und Akzeptanz. | A continuous and open dialogue with our stakeholders ensures transparency and acceptance. |
| Ein weiterer Vorteil dieser biogenen Kraftstoffe liegt darin, dass sie die Abhängigkeit von Importen fossiler Energieträger besonders effizient verringern helfen. | A further benefit of these biogenic fuels is the fact that they are extremely efficient in helping to reduce imports of fossil energy sources. |
| Ein weiterer Schwerpunkt des Risikomanagements war im vergangenen Geschäftsjahr die Einführung eines neuen konzernweit geltenden Reportingsystems zur zentralen Erfassung von vermögensschädigenden Handlungen. | A further key area of risk management in the past fiscal year was the introduction of a new Groupwide reporting system to maintain a central record of economically damaging acts. |
| Ein großer Teil dieses Warmbandes dient künftig zur Vormaterialversorgung des mexikanischen Edelstahlwerks ThyssenKrupp Mexinox. | A large portion of the hot-rolled will be supplied to the ThyssenKrupp Mexinox stainless steel plant in Mexico. |
| Dazu haben wesentlich die nordamerikanischen Aktivitäten beigetragen, die zum einen weiterhin von der guten Marktsituation profitierten und zum anderen eine höhere Margensteigerung erzielen konnten. | A major role in this was played by the North American activities which not only continued to benefit from the good market situation but also achieved higher margins. |
| Ein neuer Meilenstein in der Geschichte der Polymerschäume. | A new milestone in the history of polymer foams. |
| Bindemittelsortiment für Holzanstriche. | A range of binders for wood coatings. |
| Eine Überprüfung von Systematik und Höhe der Vorstandsvergütung erfolgt in regelmäßigen Abständen. | A review of the structure and amount of compensation of Board members takes place at regular intervals. |
| Eine Auswahl an Produkten für unser Benzin-Direkteinspritzsystem. | A selection of products for our direct gasoline injection system. |
| Eine Kleinserie dieser Solo-Fahrzeuge wird ab Ende des Jahres 2009 in mehreren europäischen Großstädten in den Linienbetrieb gehen. | A small series of these solo vehicles will be taken into regular service in several European cities from end of 2009. |
| Ein sehr vielversprechender Lösungsansatz ist die Verwendung des Generators selbst als Spannsystem. | A very promising approach is the use of the generator itself as a tensioning system. |
| Virtueller IAA-Messerundgang als App für iPhone. | A virtual tour of the IAA fair via iPhone app. |
| Ein mit Widia-Bohrstücken besetzter Fischschwanzmeißel für Tiefbohrungen, 1936. | A Widia-tipped fishtail bit for deep hole drilling, 1936. |

**Figure 1 Test sentences and reference translations for German-English pair**

For English-Croatian pair 500 sentences, with one reference translation each, were used in each of two domains.

# 4. Evaluation Methodology

## *4.1 General options*

Several methods are used for evaluation of MT results cf. Figure 2.

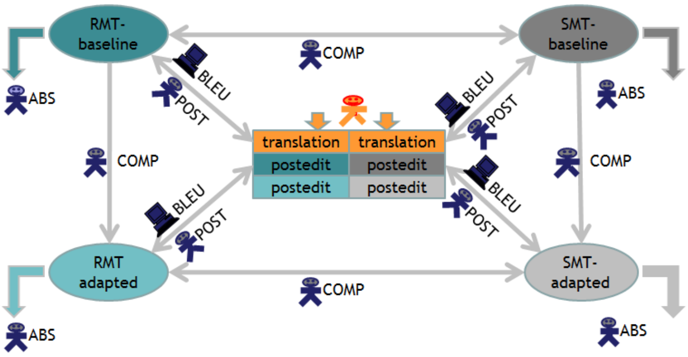


**Figure 2 Evaluation options**

**Automatic** comparison (called BLEU in Figure 2) is the predominant paradigm in the world of SMT. So BLEU and/or NIST scores can be computed for different versions of MT system output.

While such scores seem to measure inner-system quality changes with some degree of reliability, they do not seem to measure translation quality, do not conform to human evaluators' judgement, and are sensitive towards an SMT system architecture in disfavour of rule-based approaches. Therefore projects like WMT do not use them as the only measure of quality any more.

**Comparative** evaluation (called COMP in Figure 2) is relevant if two outputs of a system, or the output of two systems are to be compared. Comparative evaluation is possible between two systems as well as between two versions of the same system.

While this approach can find which of two systems has an overall better score, it cannot answer the question what the real quality of the two systems is, and what the quality baseline of the comparison looks like.

**Absolute** evaluation (called ABS in Figure 2) therefore is required to determine the quality of a given translation. This procedure looks at *one* translation of a source sentence at a time, and determines its accuracy (how much content has been transported to the target language) and fluency (how correct / grammatical is the target sentence produced?).

This evaluation method is often used in RMT technology as a threshold for a release of a language pair (e.g.: more than 70% of the test sentences must be correct or understandable in the target language). It can refer to one translation result of a time, be it baseline or adapted, of SMT or RMT output.

**Postediting** evaluation (called POST in Figure 2) reflects the task-oriented aspect of evaluation. It measures the distance of an MT output to a human (MT-postedited) output, either in terms of time (answering the question how productive a system can be as compared e.g. to a human-only translation), or in terms of the keystrokes needed to produce a human-corrected translation from an MT-raw translation (HTER).

## 4.2 Evaluation in ACCURAT

In the ACCURAT narrow domain task, the following evaluation methods were used for German-English pair: cf. Figure 3:

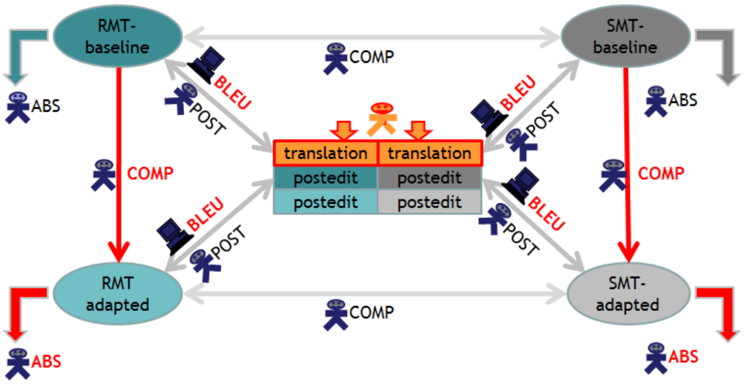


**Figure 3 Evaluation in ACCURAT**

- **Automatic** evaluation of the four systems (DFKI-baseline and DFKI-adapted, PT-baseline and PT-adapted) using BLEU and NIST scores
- **Comparative** evaluation of the pairs DFKI-baseline vs. DFKI-adapted and PT-baseline vs. PT-adapted; this would produce the core information how much the systems can improve
- **Absolute** evaluation of the systems DFKI-adapted and PT-adapted, to gain insight into the translation quality, and consequently the potential acceptance of such systems for real-world use

Other forms of evaluation (comparison between PT-adapted and DFKI-adapted, or postediting evaluation) were not included into the evaluation task, esp. postediting evaluation is done in other evaluations in the ACCURAT project.

For English-Croatian pair the following evaluation directions were used:

- **Automatic** evaluation of the four systems (enhr-automotive-baseline and enhr-automotive-adapted, enhr-REn-baseline and enhr-REn-adapted) using BLEU scores
- **Comparative** evaluation of the systems enhr-automotive-baseline vs. enhr-automotive-adapted and enhr-REn-baseline vs. enhr-REn-adapted
- **Absolute** evaluation of the systems enhr-automotive-adapted and enhr-REn-adapted.

# 5. Evaluation Tools

To perform the evaluations, a special toolset was created for the non-automatic tasks. The toolset is called ‘Sisyphos-II’, and consists of three components:

- ‘ABS’ to support absolute evaluation
- ‘COMP’ to support comparative evaluation of two MT outputs
- ‘POST’ to support postediting evaluation, by measuring the postediting time (in seconds) and allowing HTER computing

The tools are stand-alone tools which can be given e.g. to a freelance translator; evaluation data are presented to the users by a special GUI in random order, and evaluation results are collected in another XML file which is the basis for evaluation. In comparison with the versions from end-2011, they have been improved by an easier import and by an evaluation statistics component.

The documentation of Sisyphos-II is given in the annex; screenshots of the GUI are shown in Figure 4, Figure 5, and Figure 6.

The evaluation interaction differs depending on the tool. It displays sentences with their translations, in random order. Each tool has a section where the source and translations are displayed, and below that a section with the evaluation options. At the bottom of the screen, buttons for the different system possibilities are located:

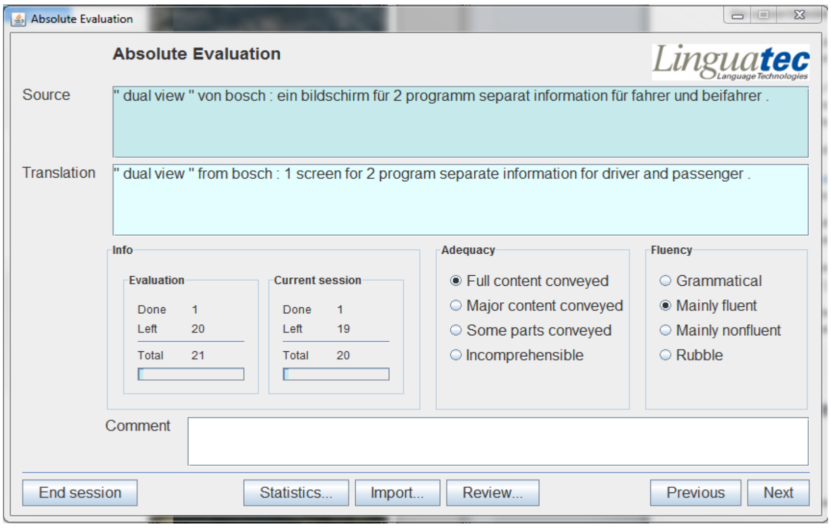
Navigation in the evaluation data is done with [Next] and [Previous]; [End Session] terminates the current session, [Import] creates a new evaluation file, [Review] accesses evaluation results of a previous session, and [Statistics] displays a table with evaluation results.

## 5.1 Absolute Evaluation

For a given translation, its quality is determined.

The translation is displayed, and users can evaluate the adequacy and the fluency of the translation. Each time a 4-point scale is presented, users select one of the options in both areas.

- For adequacy, the options are: { *full content conveyed* | *major content conveyed* | *some parts conveyed* | *incomprehensible* }
- For fluency, the options are: { *grammatical* |



Absolute Evaluation

Source: " dual view " von bosch : ein bildschirm für 2 programm separat information für fahrer und beifahrer .

Translation: " dual view " from bosch : 1 screen for 2 program separate information for driver and passenger .

Info

| Evaluation | |
|---|---|
| Done | 1 |
| Left | 20 |
| Total | 21 |

| Current session | |
|---|---|
| Done | 1 |
| Left | 19 |
| Total | 20 |

Adequacy
● Full content conveyed
○ Major content conveyed
○ Some parts conveyed
○ Incomprehensible

Fluency
○ Grammatical
● Mainly fluent
○ Mainly nonfluent
○ Rubble

Comment

**Figure 4 Absolute evaluation**

*mainly fluent* | *mainly nonfluent* | *rubble* }

By clicking on [Next] the result is stored, and the next sentence is presented, [Previous] displays previous evaluation data, for corrections.

## 5.2 Comparative evaluation

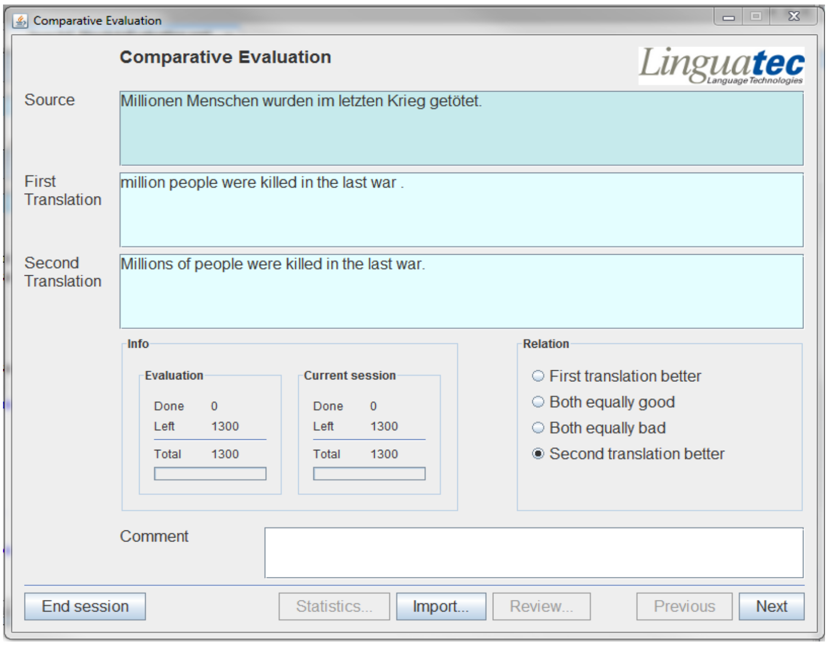
The tool compares the quality of two translations against each other.

Two translations of a given sentence are displayed, for comparison. Users can decide which one is better, on a 4-point scale.

Comparison options are: { *first translation better* | *both equally good* | *both equally bad* | *second translation better* }.

The sequence of translation1 and translation2 is randomized to avoid biased evaluation (i.e. translation 1 is sometimes displayed first, sometimes second).

By clicking on [Next] the result is stored, and the next sentence is presented, [Previous] displays previous evaluation data, for corrections.



Comparative Evaluation

Source: Millionen Menschen wurden im letzten Krieg getötet.

First Translation: million people were killed in the last war .

Second Translation: Millions of people were killed in the last war.

Info — Evaluation: Done 0, Left 1300, Total 1300; Current session: Done 0, Left 1300, Total 1300

Relation: ○ First translation better ○ Both equally good ○ Both equally bad ● Second translation better

Comment:

**Figure 5 Comparative Evaluation**

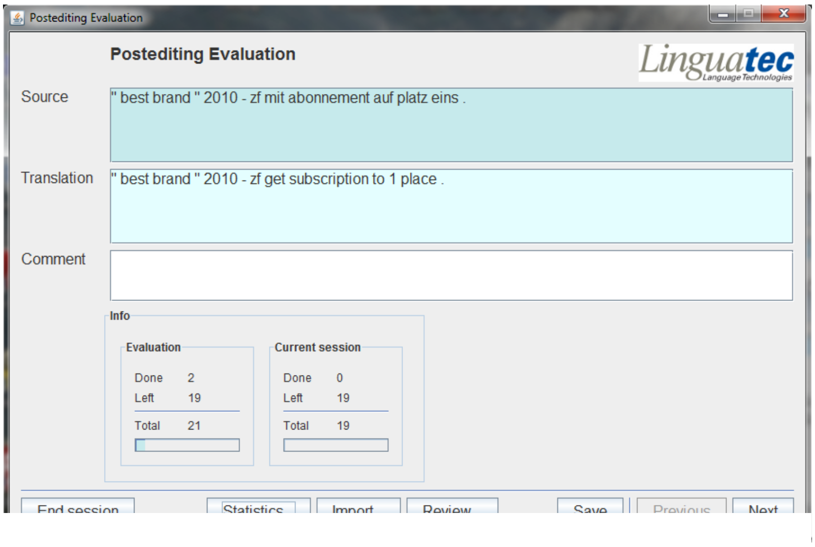
## 5.2 Postediting evaluation

The tool measures the time needed to postedit a translation output into a correct format (HTER). It can afterwards also be used to compute the edit distance.

The translation of the source sentence is displayed. The translation field is editable, so users can edit the MT output.

The time from the first display of the sentence until the pressing of the [Save] button is stored (in seconds). There is also a ‘comment’ field which can be used to give comments on the translation / postediting.

Navigation is done with the [Next] and [Previous] buttons.



Postediting Evaluation

Source: " best brand " 2010 - zf mit abonnement auf platz eins .

Translation: " best brand " 2010 - zf get subscription to 1 place .

Comment:

Info — Evaluation: Done 2, Left 19, Total 21; Current session: Done 0, Left 19, Total 19

**Figure 6 Postediting evaluation**

# 6. Evaluation Results

For German-English pair three evaluators were used to do the translations. Each of them evaluated a random subset of the 1500 sentence test set, consisting of at least 500 sentences for the COMP evaluations, and at least 300 sentences for the ABS evaluations per task. For four tasks and three evaluators, more than 5000 evaluation points were collected this way.

For English-Croatian pair four evaluators were used and each of them evaluated 125 sentences for the COMP evaluations and 125 sentences for the ABS evaluations. For two narrow domains, four evaluators collected 2000 evaluation points.

## *6.1 Automatic evaluation*

The automatic evaluation for German-English pair was done on the basis of BLEU scores. The results are:

**Table 1 Bleu scores baseline vs. adapted for German-English pair**

| | DFKI | PT |
|---|---|---|
| baseline | 17.36 | 16.08 |
| adapted | 22.21 | 17.51 |

For both systems there is an increase in BLEU; more moderate for the RMT than for the SMT system. However it is known that BLEU is biased towards SMT systems.

The automatic evaluation for English-Croatian pair was done on the basis of BLEU scores. The results are:

**Table 2 Bleu scores baseline vs. adapted for English-Croatian pair**

| | enhr-automotive | enhr-REn |
|---|---|---|
| baseline | 25.87 | 11.81 |
| adapted | 24.98 | 14.08 |

For enhr-automotive there is a decrease in BLEU score; but just less than one BLEU point. There are several possible reasons why the adaptation of the baseline system to this narrow domain didn’t produce the increase in the BLEU score, but the most probable one is the quality of data used for the adjustment of baseline model to the narrow domain. Since for two different languages texts from automotive domain were collected from different sources and with different methodology, it seems that this comparable corpus has more features of weakly comparable corpus than strongly comparable corpus. Also, the automotive domain could be understood in a very wide manner because automotive seeding terms can appear in many different areas. Since because of under-resourcedness of texts in Croatian, this approach was used for the selection of texts in Croatian.

For enhr-REn there is a significant increase of 2.27 BLUE points (19.22%) as expected with the strongly comparable corpus in the narrow domain of renewable energy and more strict ways of selecting texts for comparable corpus in both languages.

However, we wanted to check using human evaluation whether these results have counterpart in the real translation quality.

### 6.2 Comparative Evaluation

For German-English pair three testers were used, all of them good speakers of English with a bit of MT background.

Of the 1500 test sentences, three testers inspected randomly selected subsets, in total about 2000 sentences. As the tool does not offer identical sentences for evaluation, they cannot be differentiated for 'equally good' vs. 'equally bad'. If these two categories are merged into one ('equal'), the following results were achieved.

**Table 3 Comparative Evaluation baseline vs. adapted for DFKI and PT in German-English pair**

| **DFKI** | **Tester1** | **Tester2** | **Tester3** | **total** |
|---|---|---|---|---|
| base better | 235 | 130 | 82 | 447 |
| equal | 514 | 228 | 319 | 1061 |
| adaption better | 300 | 152 | 100 | 552 |
| total inspected | 1049 | 510 | 501 | 2060 |
| improvement | **6,20%** | **4,31%** | **3,59%** | **5,10%** |
| **PT** | **Tester1** | **Tester2** | **Tester3** | **total** |
| base better | 91 | 33 | 34 | 158 |
| equal | 1237 | 417 | 418 | 2072 |
| adaption better | 173 | 53 | 49 | 275 |
| total inspected | 1501 | 503 | 501 | 2505 |
| improvement | **5,46%** | **3,98%** | **2,99%** | **4,67%** |

The data show that for both types of systems, the domain adaptation results in an improvement of 5%. It is a bit more (5.1%) for the SMT than for the RMT (4.7%). The result is consistent among the testers: all of them state an improvement of the adapted versions, and all of them see a higher improvement for the SMT than for the RMT.

It may be worthwhile noticing that in the RMT evaluation, a large proportion of the test sentences (nearly 60%) came out identical in both versions, and the changes were rather small (17% of the sentences). In the SMT system, nearly no sentence came out unchanged, and the variance was between 36% and 51% (depending on the testers).

In a sideline evaluation, a comparison was made between the baseline versions of SMT and RMT, and their adapted versions.

**Table 4 Comparative Evaluation DFKI-PT for baseline and adapted systems in German-English pair**

| | **baseline** | **adapted** |
|---|---|---|
| DFKI better | 47 | 38 |
| equal | 170 | 203 |
| PT better | 284 | 260 |
| total | 501 | 489 |
| percentage | **47.3%** | **44.3%** |

The result shows, although done with only one tester, that the RMT quality is considered significantly better than the SMT quality. The main reason for this seems to be that the SMT de-en very frequently eliminates the verbs from sentences, e.g.:

*Silber wird in der Medizin seit Jahrhunderten wegen seiner antimikrobiellen Wirkung geschätzt und eingesetzt. => silver in medicine centuries for its antimicrobial effect and .*

This effect has already been observed with other SMT outputs[1].

It should be noted, however, that the distance between the systems is smaller in the adapted versions than in the baseline versions (by 3%).

For English-Croatian pair four testers were used, all of them native speakers of Croatian.

Each tester processed 125 sentences in each narrow domain.

**Table 5 Comparative Evaluation baseline vs. adapted for enhr-automotive and enhr-REn in English-Croatian pair**

| enhr-automotive | Tester1 | Tester2 | Tester3 | Tester4 | total |
|---|---|---|---|---|---|
| base better | 21 | 14 | 25 | 40 | 100 |
| equal | 95 | 98 | 94 | 83 | 370 |
| adaption better | 9 | 13 | 6 | 2 | 30 |
| total inspected | 125 | 125 | 125 | 125 | 500 |
| improvement | -9,6 | -0,8 | -15,2 | -30,4 | -14 |
| **enhr-Ren** | **Tester1** | **Tester2** | **Tester3** | **Tester4** | **total** |
| base better | 17 | 24 | 20 | 17 | 78 |
| equal | 68 | 65 | 73 | 82 | 288 |
| adaption better | 40 | 36 | 32 | 26 | 134 |
| total inspected | 125 | 125 | 125 | 125 | 500 |
| improvement | 18,4 | 9,6 | 9,6 | 7,2 | 11,2 |

The data show that for enhr-automotive the domain adaptation with selected comparable data results in a decrease of -14% which represents an important drop. The result is consistent among the testers: all of them state the deterioration in enhr-automotive adapted version, and all of them see an improvement for the enhr-REn system.

## 6.3 Absolute Evaluation

The absolute evaluation was done to find a hint how usable the resulting translation would be after the system was adapted.

For German-English pair, a total of 1100 sentences, randomly selected from the 1500 test base, were inspected by three testers. Each adequacy and fluency was measured on a scale between 1 and 4 (1 = grammatical/fully adequate, 4 = rubble/incomprehensible). Table 6 gives the result (lower average scores mean better quality):

**Table 6 Absolute evaluation for adequacy and fluency, for DFKI and PT in German-English pair**

| DFKI-adapted | Tester 1 | Tester 2 | Tester 3 | total |
|---|---|---|---|---|
| inspected | 500 | 302 | 301 | 1103 |
| | | | | |
| adequacy 1 | 89 | 52 | 59 | 200 |
| adequacy 2 | 119 | 48 | 37 | 204 |
| adequacy 3 | 284 | 156 | 77 | 517 |

[1] cf. the systems at WMT 2011.

| DFKI-adapted | Tester 1 | Tester 2 | Tester 3 | total |
|---|---|---|---|---|
| adequacy 4 | 8 | 46 | 128 | 182 |
| average | **2,42** | **2,65** | **2,91** | **2,62** |
| | | | | |
| fluency 1 | 87 | 97 | 116 | 300 |
| fluency 2 | 163 | 97 | 25 | 285 |
| fluency 3 | 238 | 93 | 31 | 362 |
| fluency 4 | 12 | 15 | 129 | 156 |
| average | **2,35** | **2,09** | **2,57** | **2,34** |
| **PT-adapted** | **Tester 1** | **Tester 2** | **Tester 3** | **total** |
| inspected | 501 | 300 | 301 | 1102 |
| | | | | |
| adequacy 1 | 210 | 106 | 149 | 465 |
| adequacy 2 | 127 | 99 | 25 | 251 |
| adequacy 3 | 150 | 80 | 55 | 285 |
| adequacy 4 | 14 | 15 | 72 | 101 |
| average | **1,94** | **2,01** | **2,17** | **2,02** |
| | | | | |
| fluency 1 | 197 | 164 | 180 | 541 |
| fluency 2 | 189 | 89 | 35 | 313 |
| fluency 3 | 100 | 42 | 34 | 176 |
| fluency 4 | 15 | 5 | 52 | 72 |
| average | **1,87** | **1,63** | **1,86** | **1,80** |

It can be seen that testers evaluate the SMT somewhat between ‘mainly’ and ‘partially’ fluent/comprehensible, and the RMT close to ‘mainly’ fluent/comprehensible. The testers agree in their evaluation, and have similar average results. The better score for the RMT may result from the ‘missing verb’ problem mentioned above.

It could be worthwhile to mention that the opinion often heard that the SMT produces more fluent output that the RMT cannot be corroborated with the evaluation data here: The RMT output is clearly considered to be more fluent than the SMT output (1.8 vs. 2.3).

An ABSolute evaluation was also done for the two baseline systems, however with one tester only. The results are given in Table 7.

**Table 7 ABS evaluation of the baseline systems in German-English pair**

| | DFKI baseline | PT baseline |
|---|---|---|
| inspected | 301 | 301 |
| | | |
| adequacy 1 | 57 | 165 |

| | DFKI baseline | PT baseline |
|---|---:|---:|
| adequacy 2 | 51 | 15 |
| adequacy 3 | 69 | 61 |
| adequacy 4 | 124 | 60 |
| average | **2,86** | **2.05** |
| | | |
| fluency 1 | 136 | 222 |
| fluency 2 | 22 | 37 |
| fluency 3 | 46 | 18 |
| fluency 4 | 97 | 24 |
| average | **2,35** | **1.48** |

The figures indicate that the system adaptation improves the accuracy of the SMT (from 2.86 baseline to 2.62 adapted), and it seems to reduce the fluency of the RMT (from 1.48 baseline to 1.80 adapted). A further error analysis would be required to find out why. The other results seem unchanged.

As far as the inter-rater agreement is concerned, the test setup made it difficult to compute it: All testers used the same test set but tested only a random subset of it. So there are only few data points common to all testers (only 20 in many cases). For those, only weak agreement could be found (with values below 0.4 in Cohen's kappa). This is shown in Table 8.

**Table 8 Inter-rater agreement (Cohen's Kappa) in German-English pair**

| System | records inspected | common datapoints | common evaluation | kappa |
|---|---:|---:|---:|---:|
| dfki-comp | 1189 | 115 | 46 | 0,38 |
| pt-comp | 1102 | 39 | 11 | 0,26 |
| | | | | |
| dfki-abs-ad | 846 | 21 | 5 | 0,22 |
| dfki-abs-fl | 846 | 21 | 4 | 0,18 |
| pt-abs-ad | 851 | 21 | 4 | 0,17 |
| pt-abs-fl | 851 | 21 | 3 | 0,11 |

Values are slightly better if more data points are available. However, all testers show consistent behaviour in the evaluation, and came to similar conclusions overall, as has been explained above.

For English-Croatian pair, a total of 500 sentences for each narrow domain were inspected by four testers. Each adequacy and fluency was measured on a scale between 1 and 4 (1 = grammatical/fully adequate, 4 = rubble/incomprehensible). Table 9 gives the result (lower average scores mean better quality).

**Table 9 Absolute evaluation for adequacy and fluency, for enhr-automotive and enhr-REn in English-Croatian pair**

| **enhr-automotive** | **Tester 1** | **Tester 2** | **Tester 3** | **Tester 4** | **total** |
|---|---|---|---|---|---|
| inspected | 125 | 125 | 125 | 125 | 500 |
| | | | | | |
| adequacy 1 | 23 | 18 | 5 | 1 | 47 |
| adequacy 2 | 50 | 97 | 21 | 6 | 174 |
| adequacy 3 | 50 | 10 | 57 | 54 | 171 |
| adequacy 4 | 2 | 0 | 42 | 64 | 108 |
| average | 1,25 | 0,94 | 2,09 | 2,45 | **1,68** |
| | | | | | |
| fluency 1 | 23 | 22 | 1 | 0 | 46 |
| fluency 2 | 32 | 67 | 11 | 4 | 114 |
| fluency 3 | 53 | 34 | 37 | 16 | 140 |
| fluency 4 | 17 | 2 | 76 | 105 | 200 |
| average | 1,51 | 1,13 | 2,5 | 2,81 | 1,99 |
| **enhr-REn** | **Tester 1** | **Tester 2** | **Tester 3** | **Tester 4** | **total** |
| inspected | 125 | 125 | 125 | 125 | 500 |
| | | | | | |
| adequacy 1 | 24 | 15 | 32 | 40 | 111 |
| adequacy 2 | 41 | 59 | 81 | 72 | 253 |
| adequacy 3 | 55 | 48 | 12 | 12 | 127 |
| adequacy 4 | 5 | 3 | 0 | 1 | 9 |
| average | 1,33 | 1,31 | 0,84 | 0,79 | 1,07 |
| | | | | | |
| fluency 1 | 6 | 4 | 1 | 2 | 13 |
| fluency 2 | 15 | 18 | 98 | 97 | 228 |
| fluency 3 | 48 | 53 | 26 | 25 | 152 |
| fluency 4 | 56 | 50 | 0 | 1 | 107 |
| average | 2,23 | 2,19 | 1,2 | 1,2 | 1,71 |

It can be seen that testers evaluate the enhr-automotive somewhere between ‘mainly’ and ‘partially’ fluent/comprehensible, and the enhr-REn close to ‘mainly’ fluent/comprehensible. However, the testers had significant discrepancy in their evaluation marks in enhr-automotive (0.94 – 2.45), probably also showing the differences in collected texts. In the same time the adequacy for enhr-REn exhibits the best score. The better overall score for the enhr-REn may be result of the difference in strongly comparable data used for adaptation of SMT system mentioned above, reflecting them to the quality of adapted SMT system.

**An ABSolute evaluation was also done for the two English-Croatian baseline systems, however with one however with one tester only. The results are given in**

Table 10.

**Table 10 ABS evaluation of the baseline systems in English-Croatian pair**

| | enhr-automotive | enhr-REn |
|---|---|---|
| inspected | 125 | 125 |
| | | |
| adequacy  1 | 22 | 43 |
| adequacy 2 | 57 | 42 |
| adequacy 3 | 41 | 34 |
| adequacy 4 | 5 | 6 |
| average | 1,23 | 1,02 |
| | | |
| fluency 1 | 1 | 10 |
| fluency 2 | 42 | 46 |
| fluency 3 | 65 | 56 |
| fluency 4 | 17 | 13 |
| average | 1,78 | 1,58 |

The figures indicate that the system adaptation deteriorated the accuracy of the enhr-automotive system only slightly (from 1.23 baseline to 1.68 adapted). Just a bit worse, but still similar figures are found in enhr-REn system (from 1.58 baseline to 1.71 adapted). A further error analysis would be required to find out why, particularly since the BLEU scores differ much and their direction is also different.

Also, the overall averages for English-Croatian are somewhat better than overall averages for German-English pair, but we believe these are incomparable since different sets of testers were used for different language pairs and with different individual experiences in evaluating MT output. So the overall averages should be considered separate for a certain language pair.

# 7 Conclusion

For German-English pair, Figure 7 gives all evaluation results.

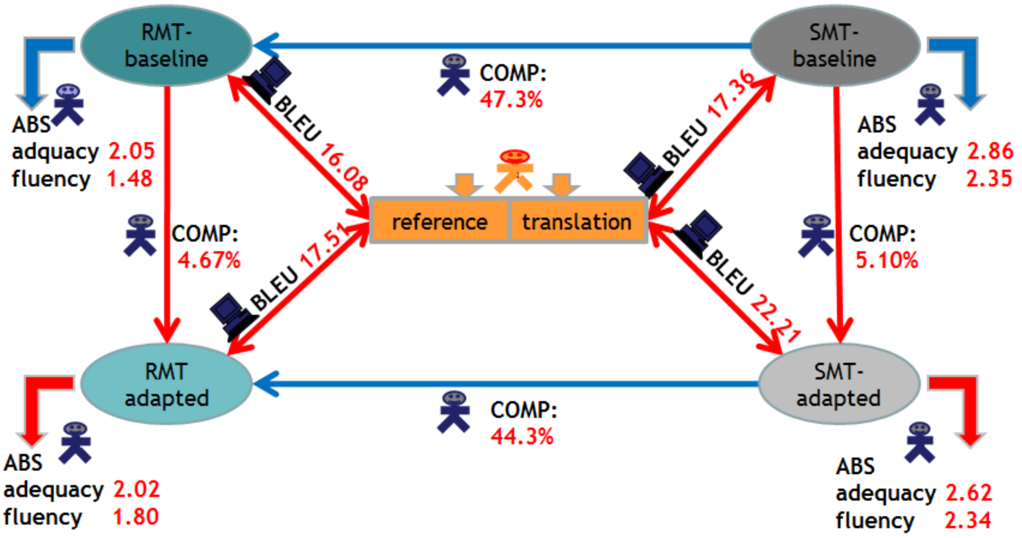


**Figure 7 Evaluation result summary for German-English pair**

The main conclusion is that all evaluation methods indicate an improvement of the adapted versions over the baseline versions:

- automatic evaluation:
  - For SMT, the BLEU score increases from 17.36 to 22.21
  - For RMT, the BLEU score increases from 16.08 to 17.51
- comparative evaluation:
  - For SMT, an improvement of 5.1% was found
  - For RMT, and improvement of 4.67% was found
- absolute evaluation:
  - For SMT, adequacy improved from 2.86 to 2.62, and fluency slightly from 2.35 to 2.34
  - For RMT, adequacy improved from 2.05 to 2.02, only fluency decreased from 1.48 to 1.8

The improvement is more significant for the SMT system than for the RMT; this may be due to the fact that the RMT baseline system has better COMP and ABS scores, although lower BLEU scores, than the SMT baseline.

For SMT improvement, (Pecina et al. 2012) report improvements between 8.6 and 16.8 BLEU for domain adaptation; results here may be a bit lower maybe due to difference in language, and a still significant percentage of OOV words.

For English-Croatian pair fig. 8 gives all results.

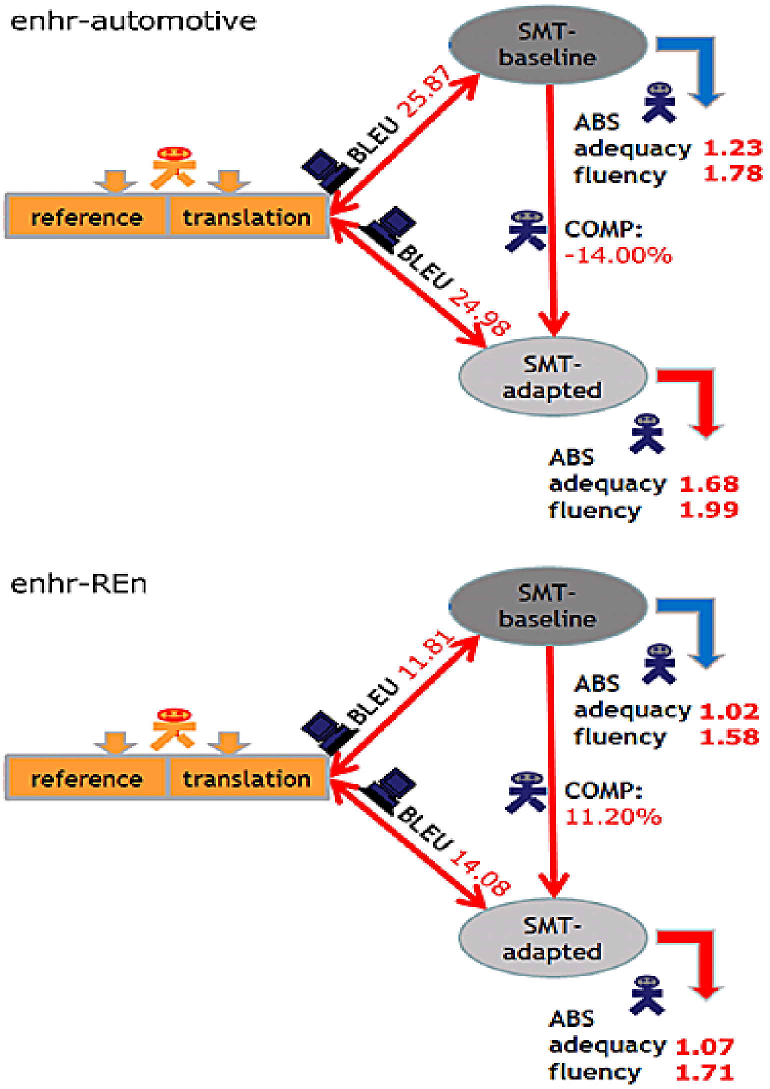


**Figure 8 Evaluation result summary for English-Croatian pair**

The main conclusion is that in some evaluation methods an improvement of the adapted versions over the baseline versions can be detected, either in improved BLEU score or adequacy:

- automatic evaluation (SMT):
  - For enhr-automotive, the BLEU score decreases slightly from 25.87 to 24.98
  - For enhr-REn, the BLEU score increases from 11.81 to 14.08
- comparative evaluation:
  - For enhr-automotive adapted, a deterioration of -14.00% was found
  - For enhr-REn adapted, an improvement of 11.20% was found
- absolute evaluation:
  - For enhr-automotive, adequacy deteriorated from 1.23 to 1.68, and fluency from 1.78 to 1.98
  - For enhr-REn, adequacy deteriorated from 1.02 to 1.07, and fluency decreased from 1.58 to 1.71

The deterioration is more significant for the enhr-automotive system while improvement is present for the enhr-REn system, and this may be due to the fact that the enhr-REn system has better COMP and ABS scores, although lower BLEU scores, than the enhr-automotive system.

# Literature

- Aleksić, V., Thurmair, Gr., 2011: Personal Translator at WMT 2011 // Proceedings of the WMT Edinburgh, UK.
- Ljubešić, N., Erjavec, T. 2011: *hrWaC and slWac:* Compiling *Web Corpora for Croatian and Slovene* // Proceedings of the 14th International Conference Text, Speech and Dialogue (TSD2011), Plzeň, Czech Republic, 2011, Lecture Notes in Artificial Intelligence 6836, Springer, Heidelberg, pp 395-402.
- Pecina, P., Toral, A., Papavassiliou, V., Prokopidis, P., van Genabith, J., 2012: Domain Adaptation of Statistical machine Translation using Web-Crawled Resources: A Case Study // Proceedings of the EAMT2012, Trento, Italy.
- Thurmair, Gr., Aleksić, V., 2012: Creating Term and Lexicon Entries from Phrase Tables // Proceedings of the EAMT2012 Trento, Italy.

# Annex: Sisyphos-II MT-Evaluation tools

D. Kaumans, Gr. Thurmair, Linguatec
2012-04-27

## *1 Introduction*

This is a set of tools for the evaluation of MT output interactively[2]. It supports the main non-automatic evaluation metrics used today, which is:

- Determination of the quality of an MT output, in terms of adequacy and fluency (called 'absolute evaluation'). This answers the question 'How good is the MT output'.
- Determination of the quality of an MT output in comparison to another MT output (called 'comparative evaluation'). It answers the question 'Which output (of two systems) is better?'. Note that it does not answer the question on the real output quality.
- Determination of the distance of an MT output to a correct human translation (called 'postediting evaluation'). It answers the question on the effort needed to create a good translation from a raw MT output, both in terms of edit distance and of required postediting time.

Three little standalone tools have been created to support these evaluations; they can be given to external evaluators (freelancers etc.), together with a pack of evaluation data, so evaluators can process them offline, and return the results. This workflow can be seen as an alternative to online-access tools as used in WMT.

## *2 Installation*

Installation requirement is a Java runtime (1.7 and higher).

The tools are deployed in a zip file. They must be extracted into a directory of users' choice; this directory will contain both the programme and the files used for processing. Below this working directory there is a directory 'lib' containing an auxiliary jar-file (for XML code handling).

The programmes are called:

- AbsoluteEvaluation.jar
- ComparativeEvaluation.jar
- PostEditingEvaluation.jar

The installation package also contains three example files, for easier startup, and the DTDs for the evaluation files.

It also contains this documentation.

## *3 Functionality*

The main functionality of the tools is:

- Import of a new evaluation 'package'
- Interactive support of the evaluation procedure
- Creation of result files containing statistics.

[2] The first version of Sisyphus was created by the Belgian METAL team in 1987, in pre-Windows times, to speed up system development. The kind of tools is still needed…

The **data flow** is depicted in Figure 1A. The main files are the translation and evaluation xml files. Each tool works with two XML files, called ‘translation-{abs|comp|post}.xml’ (created by the import function from the source and target language files produced by the MT systems), storing the data to be evaluated, and ‘evaluation-{abs|comp|post}.xml’, created during interactive evaluation, storing the evaluation result. The file names are fixed. The result of the evaluation is stored in the evaluation xml files; an overview file can be created containing basic statistics.

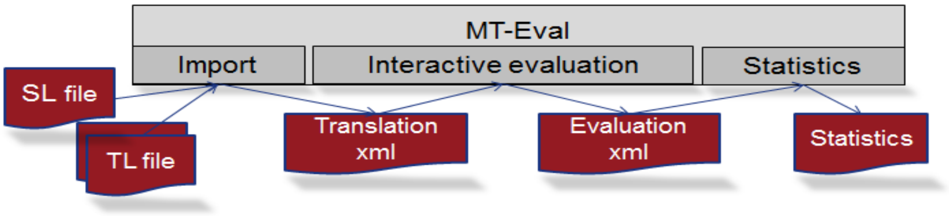


**Figure 1A Data flow**

## 3.1 Import of evaluation data

The tool expects the evaluation data in the following format:

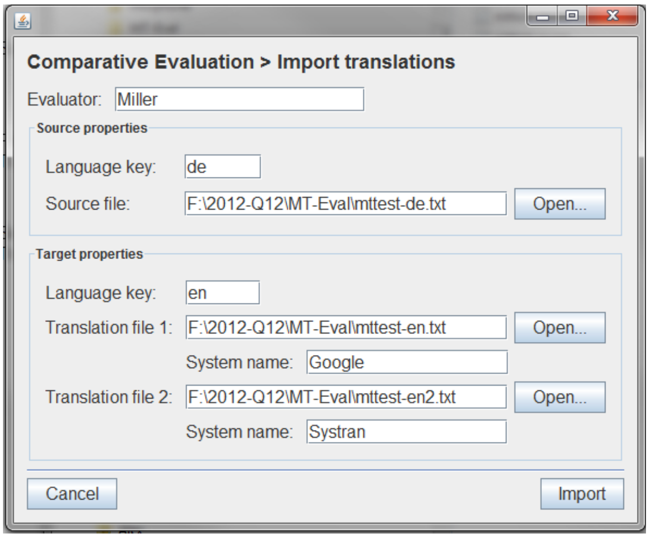
- UTF8 character code
- one line per sentence
- one file per language
- parallel numbering of sentences.

This is the basic format as produced by systems like MOSES.

By clicking on [Import] in one of the tools, the import screen is displayed, asking for

- The name / id of the evaluator
- Source and target language involved
- File name of the source and the target language(s) file
- Source of translation (which system did the translation)

With this information, an XML file is created which is used during the evaluation process. Its name is ‘translations-{comp|abs|post}.xml’ (depending on the tool). This file is used as input by the interactive evaluation process.



Comparative Evaluation > Import translations

Evaluator: Miller

Source properties

Language key: de

Source file: F:\2012-Q12\MT-Eval\mttest-de.txt

Target properties

Language key: en

Translation file 1: F:\2012-Q12\MT-Eval\mttest-en.txt

System name: Google

Translation file 2: F:\2012-Q12\MT-Eval\mttest-en2.txt

System name: Systran

**Figure 2A Data import**

## 3.2 Interactive Evaluation

The evaluation interaction differs depending on the tool. It displays sentences with their translations, in random order. Each tool has a section where the source and translations are displayed, and below that a section with the evaluation options. At the bottom of the screen, buttons for the different system possibilities are located:

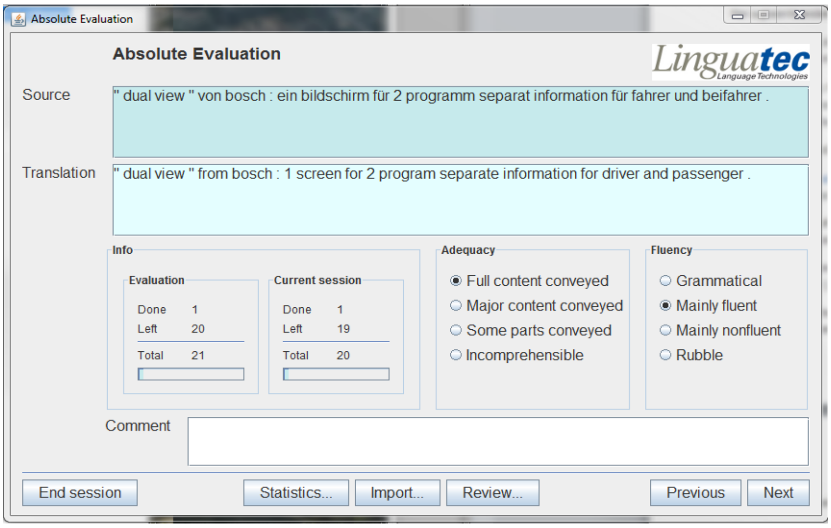
Navigation in the evaluation data is done with [Next] and [Previous]; [End Session] terminates the current session, [Import] creates a new evaluation file, [Review] accesses evaluation results of a previous session, and [Statistics] displays a table with evaluation results.

### 3.2.1 Absolute Evaluation

For a given translation, its quality is determined.

The translation is displayed, and users can evaluate the adequacy and the fluency of the translation. Each time a 4-point scale is presented, users select one of the options in both areas.

- For adequacy, the options are: { *full content conveyed | major content conveyed | some parts conveyed | incomprehensible* }
- For fluency, the options are: { *grammatical | mainly fluent | mainly nonfluent | rubble* }



Absolute Evaluation

Source: " dual view " von bosch : ein bildschirm für 2 programm separat information für fahrer und beifahrer .

Translation: " dual view " from bosch : 1 screen for 2 program separate information for driver and passenger .

Info

| Evaluation | | Current session | |
|---|---|---|---|
| Done | 1 | Done | 1 |
| Left | 20 | Left | 19 |
| Total | 21 | Total | 20 |

Adequacy
● Full content conveyed
○ Major content conveyed
○ Some parts conveyed
○ Incomprehensible

Fluency
○ Grammatical
● Mainly fluent
○ Mainly nonfluent
○ Rubble

Comment

**Figure 3A Absolute evaluation**

By clicking on [Next] the result is stored, and the next sentence is presented, [Previous] displays previous evaluation data, for corrections.

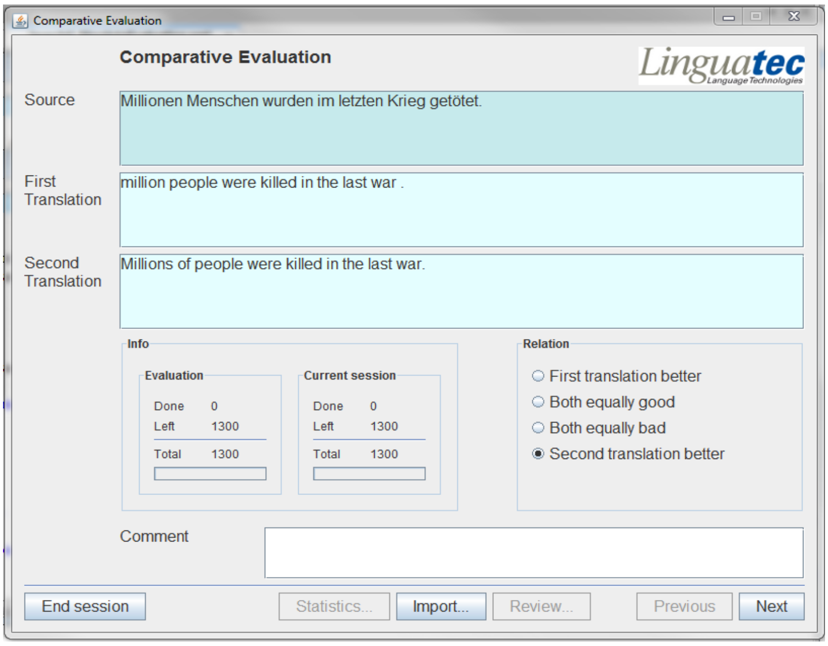
### 3.2.2 Comparative evaluation

The tool compares the quality of two translations against each other.

Two translations of a given sentence are displayed, for comparison. Users can decide which one is better, on a 4-point scale.

Comparison options are: { *first translation better | both equally good | both equally bad | second translation better* }.

The sequence of translation1 and translation2 is randomized to avoid biased evaluation (i.e. translation 1 is sometimes displayed first, sometimes second).



Comparative Evaluation

Source: Millionen Menschen wurden im letzten Krieg getötet.

First Translation: million people were killed in the last war .

Second Translation: Millions of people were killed in the last war.

Info

| Evaluation | | Current session | |
|---|---|---|---|
| Done | 0 | Done | 0 |
| Left | 1300 | Left | 1300 |
| Total | 1300 | Total | 1300 |

Relation
○ First translation better
○ Both equally good
○ Both equally bad
● Second translation better

Comment

**Figure 4A Comparative Evaluation**

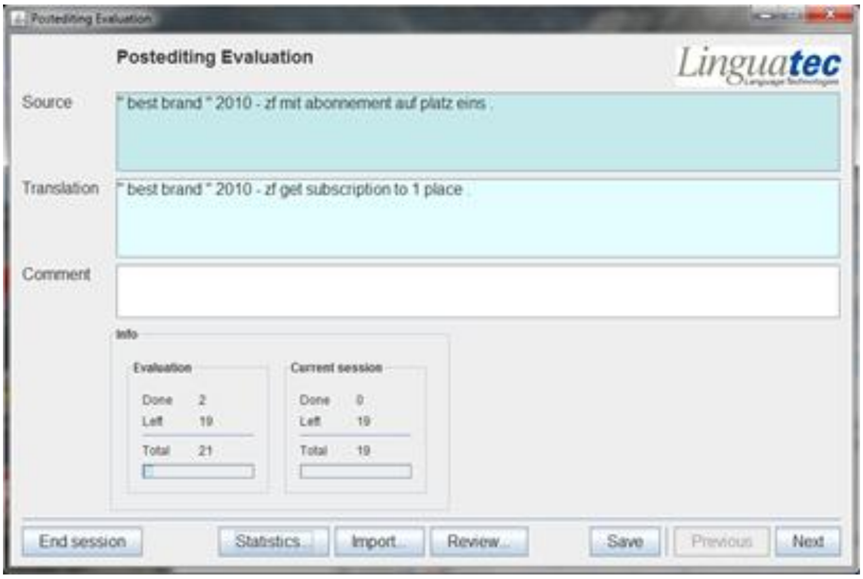
By clicking on [Next] the result is stored, and the next sentence is presented, [Previous] displays previous evaluation data, for corrections.

### 3.2.3 Postediting evaluation

The tool measures the time needed to postedit a translation output into a correct format (HTER). It can afterwards also be used to compute the edit distance.

The translation of the source sentence is displayed. The translation field is editable, so users can edit the MT output.

The time from the first display of the sentence until the pressing of the [Save] button is stored (in seconds). There is also a ‘comment’ field which can be used to give comments on the translation / postediting. Navigation is done with the [Next] and [Previous] buttons.



Posteditng Evaluation

Source: " best brand " 2010 - zf mit abonnement auf platz eins .

Translation: " best brand " 2010 - zf get subscription to 1 place .

Evaluation: Done 2, Left 19, Total 21

Current session: Done 0, Left 19, Total 19

**Figure 5A Postediting evaluation**

### 3.2.4 Common features

All tools have common features; this relates mainly to the concepts of sessions. Usually people cannot do the complete evaluation in one go, but do it in several sessions.

Within a session, users can move back and forth in the evaluated sentences, and also go back and correct an evaluation, by clicking on [Previous]. Also, a statistics on the progress of the current session is displayed, as well as of the whole task. This is for motivation reasons. If users want to stop they click on [End session].

If a session is closed, another XML file containing the evaluation results is written / updated. This file is called evaluation-{abs|comp|post}.xml.

Users can also access the evaluations of a previous session by clicking on [Review]. This allows them to change evaluation results from previous sessions (i.e. modify the evaluation-xml file). The system displays the evaluated sentence pairs, users can click on the one they want to change, and click on [edit] to edit it. This is relevant as sometimes the evaluation criteria change after having seen the first couple of data.

## 3.3 Evaluation

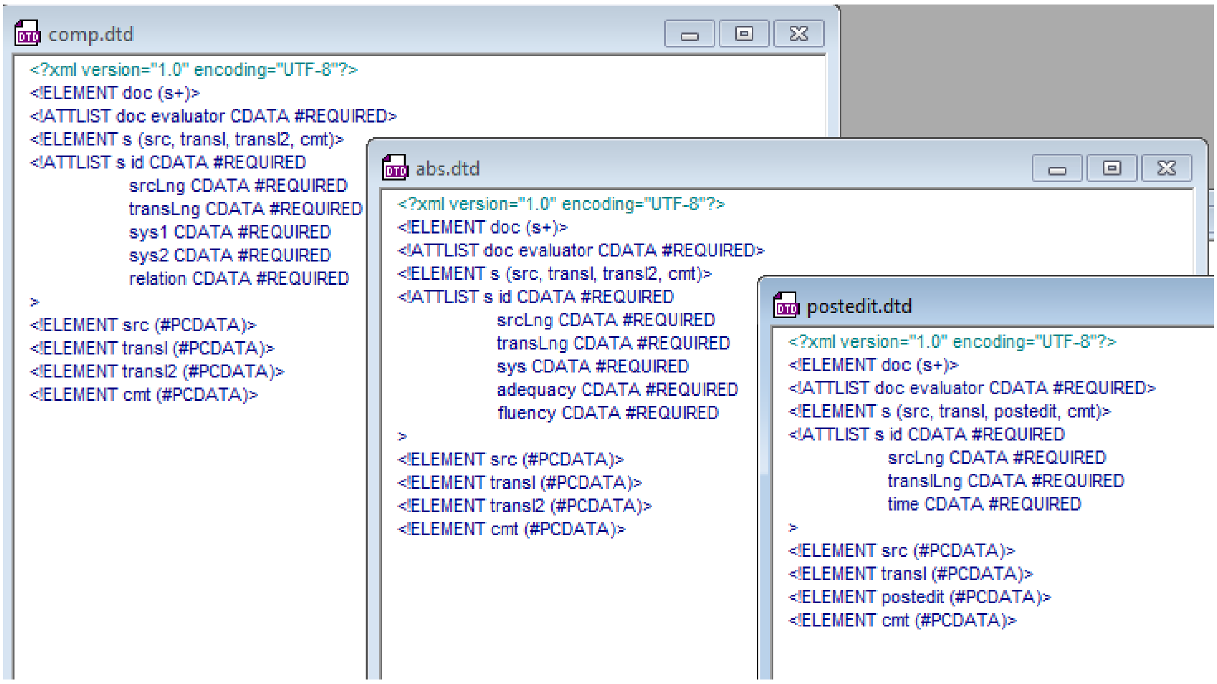
Users have the option to see an overview of the evaluation at any time of their work. They can click on [Statistics], and then a first statistics on the number of sentences, and how they were evaluated, is shown. Users can print this into a file.

For more detailed evaluation, the evaluation XML files used by the tools must be consulted, like for inter-annotator agreement, or for edit-distance computation. The format of the different tools differs slightly; the DTD of them is given in Figure 6A. Examples of the files are given in Figure 7A (for easier processing, all XML markups are in separate lines).

From this XML file, the interesting data can be extracted, e.g.:

- for Kappa calculation: sentence IDs, evaluator, evaluation results
- for edit distance calculation: translated text vs. postedited text, etc.

Users should save away the evaluation XML files from the working directory of the MT-Eval tools, to protect them from being overwritten by the next evaluation task.

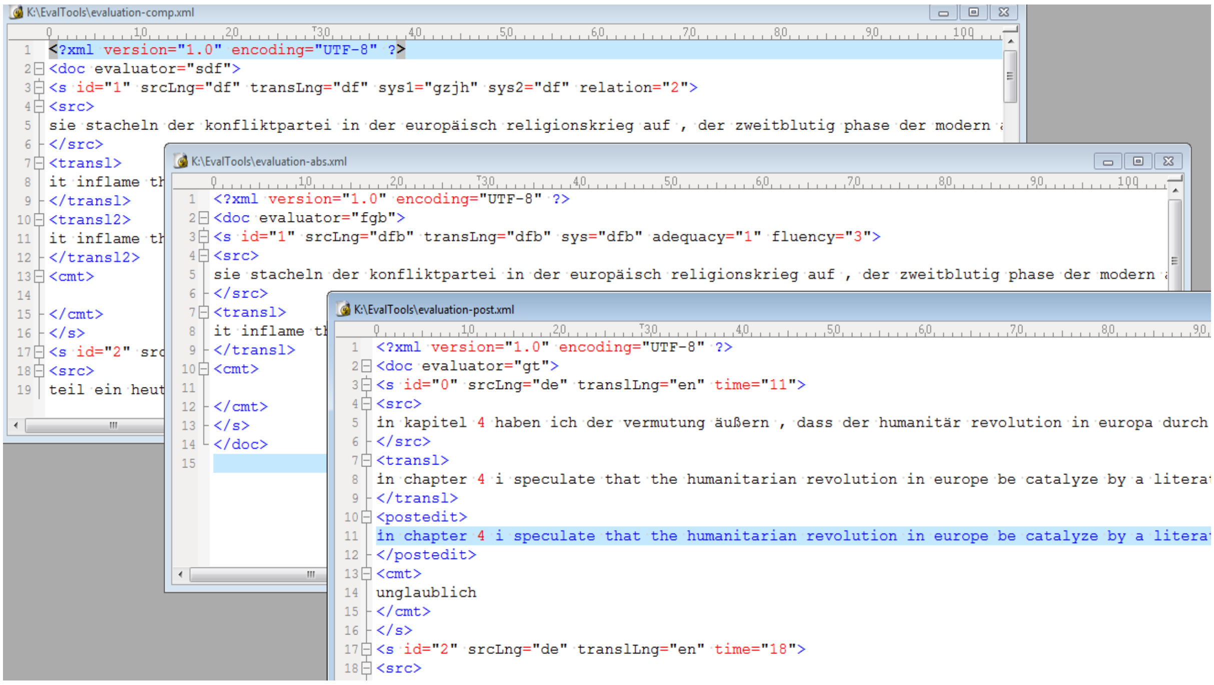


```
<?xml version="1.0" encoding="UTF-8"?>
<!ELEMENT doc (s+)>
<!ATTLIST doc evaluator CDATA #REQUIRED>
<!ELEMENT s (src, transl, transl2, cmt)>
<!ATTLIST s id CDATA #REQUIRED
            srcLng CDATA #REQUIRED
            transLng CDATA #REQUIRED
            sys1 CDATA #REQUIRED
            sys2 CDATA #REQUIRED
            relation CDATA #REQUIRED
>
<!ELEMENT src (#PCDATA)>
<!ELEMENT transl (#PCDATA)>
<!ELEMENT transl2 (#PCDATA)>
<!ELEMENT cmt (#PCDATA)>
```

```
<?xml version="1.0" encoding="UTF-8"?>
<!ELEMENT doc (s+)>
<!ATTLIST doc evaluator CDATA #REQUIRED>
<!ELEMENT s (src, transl, transl2, cmt)>
<!ATTLIST s id CDATA #REQUIRED
            srcLng CDATA #REQUIRED
            transLng CDATA #REQUIRED
            sys CDATA #REQUIRED
            adequacy CDATA #REQUIRED
            fluency CDATA #REQUIRED
>
<!ELEMENT src (#PCDATA)>
<!ELEMENT transl (#PCDATA)>
<!ELEMENT transl2 (#PCDATA)>
<!ELEMENT cmt (#PCDATA)>
```

```
<?xml version="1.0" encoding="UTF-8"?>
<!ELEMENT doc (s+)>
<!ATTLIST doc evaluator CDATA #REQUIRED>
<!ELEMENT s (src, transl, postedit, cmt)>
<!ATTLIST s id CDATA #REQUIRED
            srcLng CDATA #REQUIRED
            translLng CDATA #REQUIRED
            time CDATA #REQUIRED
>
<!ELEMENT src (#PCDATA)>
<!ELEMENT transl (#PCDATA)>
<!ELEMENT postedit (#PCDATA)>
<!ELEMENT cmt (#PCDATA)>
```

**Figure 6A DTDs of evaluation files**


K:\EvalTools\evaluation-comp.xml
K:\EvalTools\evaluation-abs.xml
K:\EvalTools\evaluation-post.xml

```
<?xml version="1.0" encoding="UTF-8" ?>
<doc evaluator="sdf">
<s id="1" srcLng="df" transLng="df" sys1="gzjh" sys2="df" relation="2">
<src>
sie stacheln der konfliktpartei in der europäisch religionskrieg auf , der zweitblutig phase der modern
</src>
<transl>
it inflame th
</transl>
<transl2>
it inflame th
</transl2>
<cmt>

</cmt>
</s>
<s id="2" src
<src>
teil ein heut
```

```
<?xml version="1.0" encoding="UTF-8" ?>
<doc evaluator="fgb">
<s id="1" srcLng="dfb" transLng="dfb" sys="dfb" adequacy="1" fluency="3">
<src>
sie stacheln der konfliktpartei in der europäisch religionskrieg auf , der zweitblutig phase der modern
</src>
<transl>
it inflame th
</transl>
<cmt>

</cmt>
</s>
</doc>
```

```
<?xml version="1.0" encoding="UTF-8" ?>
<doc evaluator="gt">
<s id="0" srcLng="de" translLng="en" time="11">
<src>
in kapitel 4 haben ich der vermutung äußern , dass der humanitär revolution in europa durch
</src>
<transl>
in chapter 4 i speculate that the humanitarian revolution in europe be catalyze by a litera
</transl>
<postedit>
in chapter 4 i speculate that the humanitarian revolution in europe be catalyze by a litera
</postedit>
<cmt>
unglaublich
</cmt>
</s>
<s id="2" srcLng="de" translLng="en" time="18">
<src>
```

**Figure 7A Example of Evaluation Files**