Machine Translation Enhanced Computer Assisted Translation

D5.2 - Evaluation Plan

Authors: Alessandro Cattelan, Mauro Cettolo
Dissemination Level: Public
Date: 30 April 2012
## D5.2 – Evaluation Plan

<table>
<thead>
<tr>
<th>Grant agreement no.</th>
<th>287688</th>
</tr>
</thead>
<tbody>
<tr>
<td>Project acronym</td>
<td>MateCat</td>
</tr>
<tr>
<td>Project full title</td>
<td>Machine Translation Enhanced Computer Assisted Translation</td>
</tr>
<tr>
<td>Funding scheme</td>
<td>Collaborative project</td>
</tr>
<tr>
<td>Coordinator</td>
<td>Marcello Federico (FBK)</td>
</tr>
<tr>
<td>Start date, duration</td>
<td>November 1st 2011, 36 months</td>
</tr>
<tr>
<td>Dissemination level</td>
<td>Public</td>
</tr>
<tr>
<td>Contractual date of delivery</td>
<td>April 30, 2012</td>
</tr>
<tr>
<td>Actual date of delivery</td>
<td>May 31, 2012</td>
</tr>
<tr>
<td>Deliverable number</td>
<td>5.2</td>
</tr>
<tr>
<td>Deliverable title</td>
<td>Evaluation Plan</td>
</tr>
<tr>
<td>Type</td>
<td>Report</td>
</tr>
<tr>
<td>Status and version</td>
<td>Final, V1.2</td>
</tr>
<tr>
<td>Number of pages</td>
<td>26</td>
</tr>
<tr>
<td>Contributing partners</td>
<td>TRANSLATED, FBK, LE MANS, UEDIN</td>
</tr>
<tr>
<td>WP leader</td>
<td>Translated</td>
</tr>
<tr>
<td>Task leader</td>
<td>Translated</td>
</tr>
<tr>
<td>Authors</td>
<td>Alessandro Cattelan, Mauro Cettolo</td>
</tr>
<tr>
<td>Reviewers</td>
<td>Loïc Barrault</td>
</tr>
<tr>
<td>EC project officer</td>
<td>Alexandra Wesolowska</td>
</tr>
<tr>
<td>The partners in MateCat are:</td>
<td>Fondazione Bruno Kessler (FBK), Italy</td>
</tr>
<tr>
<td></td>
<td>Université Le Mans (LE MANS), France</td>
</tr>
<tr>
<td></td>
<td>The University of Edinburgh (UEDIN)</td>
</tr>
<tr>
<td></td>
<td>Translated S.r.l. (TRANSLATED)</td>
</tr>
</tbody>
</table>

For copies of reports, updates on project activities and other MateCat-related information, contact:

**FBK**
MateCat
Povo - Via Sommarive 18
I-38123 Trento, Italy
Phone: +39 0461 314 521
Fax: +39 0461 314 591

© 2012, Alessandro Cattelan, Mauro Cettolo
No part of this document may be reproduced or transmitted in any form, or by any means, electronic or mechanical, including photocopy, recording, or any information storage and retrieval system, without permission from the copyright owner.
Executive Summary

This document defines evaluation methodologies and benchmarks for laboratory testing of self-tuning MT, user-adaptive MT, informative MT, and field-testing of the MateCat tool. After a short introduction, which includes the scheduling of planned evaluations, a brief description and statistics of corpora acquired for performing experiments are provided. Then, the metric for assessing the quality of MT models evolving over time is discussed in some details, as it is novel with respect to the state of the art. It will be exploited in lab tests, together with standard metric scores. The document ends with the discussion of methodologies for the evaluation of field tests.
Table of Content

Executive Summary ........................................................................................................... 3
Table of Content ................................................................................................................ 4
1 Introduction .................................................................................................................... 5
2 In-Domain Training Data ................................................................................................. 6
   2.1 Legal Domain ........................................................................................................... 6
   2.3 IT Domain .............................................................................................................. 6
3 Benchmarks for Lab Tests ............................................................................................... 7
   3.1 Legal Domain ........................................................................................................... 7
   3.2 IT Domain .............................................................................................................. 8
4 Benchmarks for Field Tests ............................................................................................ 8
   4.1 Legal Domain ........................................................................................................... 9
   4.2 IT Domain .............................................................................................................. 9
5 Evaluation Methodologies ............................................................................................... 9
   5.1 Lab Tests ................................................................................................................. 9
       5.1.1 Dynamic Adaptation Framework ................................................................. 12
       5.1.2 The Percentage Slope ................................................................................. 13
       5.1.3 Experimental assessment of the Percentage Slope .................................... 14
       5.1.4 Open Issues ................................................................................................. 16
   5.2 Field Tests .............................................................................................................. 17
       5.2.1 Productivity test ............................................................................................... 18
       5.2.2 Contrastive test ............................................................................................... 20
       5.2.3 Usability study ................................................................................................. 21
6 Appendix 1. Draft of usability questionnaire .................................................................... 24
7 Bibliography ..................................................................................................................... 26
1 Introduction

In MateCat, by its nature, evaluations play a key role as through them the effectiveness of specific design/implementation choices can be continuously assessed and the progress of the MateCat tool measured.

Evaluations are of two types: those performed in laboratory by academic partners during the daily development of models and those performed on the field by Translated involving professional translators. In lab tests, performance of system components will be evaluated, while in field tests user productivity with the MateCat tool will be measured. Lab and field tests will be conducted according to the life cycle of the project summarized in Figure 1 and described in in Section B.1.3.1 of the Description of Work.

Four evaluations will be carried out on improved versions of the MateCat tool through field tests. The analysis of the outcomes of each evaluation will provide feedback to the development of the following versions. Measures listed in Section “Evaluation Methodologies - Field Tests” will be taken; they have been chosen with the goal of making feedback as much informative as possible and will be performed on the benchmarks described in Section “Benchmarks for Field Tests”.

Single components of the MateCat tool, mainly the MT sub-modules, will be continuously evaluated during their development by means of lab tests. In addition to the standard objective metrics commonly employed by the MT research community, we will adopt novel measures specifically designed for assessing the new MT functionalities, namely domain and project adaptation in self-tuning MT, on-line learning and context-aware translation in user-adaptive MT. Evaluation methodologies for lab tests are described in Section “Evaluation Methodologies - Lab Tests”, while corresponding benchmarks are presented in “Section Benchmarks for Lab Tests”.

Simple evaluation methods have also to be defined to assess the general requirements of the MateCat tool, in terms of software reliability and response time.

Figure 1 – Project life cycle (figure taken from the Description of Work, section B1.3.1).
2 In-Domain Training Data

2.1 Legal Domain

In the Legal domain, texts from publicly available corpora have been collected:

- **JRC-Acquis**: it is the version of the Acquis Communautaire (AC) released by JRC for research purposes. AC is the total body of European Union law applicable in the EU Member States. Sentence-level alignments were generated with three tools, namely Vanilla, HunAlign and Gargantua. A preliminary investigation on their quality yielded to select the Gargantua alignment for MT experiments.
- **DGT-TM**: it is the multilingual Translation Memory for the AC made available by the European Commission’s Directorate-General for Translation.
- **OPUS-EUconst**: it is the European Constitution included in the OPUS collection.
- **OPUS-ECB**: it is the European Central Bank corpus included in the OPUS collection.

### Table 1 – Statistics of parallel training data available for the Legal Domain.

<table>
<thead>
<tr>
<th></th>
<th>en-de</th>
<th>en-es</th>
<th>en-fr</th>
<th>en-it</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>#sent (k)</td>
<td></td>
<td>#sent (k)</td>
<td></td>
</tr>
<tr>
<td><strong>ACQUIS</strong></td>
<td>2.7M</td>
<td>53.6</td>
<td>49.3</td>
<td>2.7M</td>
</tr>
<tr>
<td><strong>DGT-TM</strong></td>
<td>531</td>
<td>9.3</td>
<td>10.5</td>
<td>507</td>
</tr>
<tr>
<td><strong>OPUS</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>ECB</strong></td>
<td>113</td>
<td>3.1</td>
<td>2.8</td>
<td>116</td>
</tr>
</tbody>
</table>

2.3 IT Domain

In the IT domain, collected texts are from:

- **Translated TMs**: a number of translation memories provided by Translated.
- **OPUS-KDEdoc**: it is the manual of KDE. KDE is a Windowing Manager and Graphical User Interface for the UNIX operating system.
- **OPUS-KDE4**: it is the parallel corpus of localization files of KDE4, the current version of KDE.
- **OPUS-Open Office**: a collection of documents related to Open Office, a free and open productivity suite.
- **OPUS-Open Office 3**: another version (v3) of Open Office documentation.
Machine Translation Enhanced Computer Assisted Translation

D5.2 – Evaluation Plan

- OPUS-PHP: it is the manual corpus of PHP, a scripting language for Web development.

Table 2 – Statistics of parallel training data available for the IT domain.

<table>
<thead>
<tr>
<th></th>
<th>en-de</th>
<th>en-es</th>
<th>en-fr</th>
<th>en-it</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>#sen</td>
<td>src</td>
<td>tgt</td>
<td>#sen</td>
</tr>
<tr>
<td></td>
<td>(k)</td>
<td>(M)</td>
<td>(M)</td>
<td>(k)</td>
</tr>
<tr>
<td>Translated-TM</td>
<td>147</td>
<td>2.4</td>
<td>2.3</td>
<td>210</td>
</tr>
<tr>
<td>KDE4-GB</td>
<td>176</td>
<td>1.1</td>
<td>1.0</td>
<td>175</td>
</tr>
<tr>
<td>KDE4</td>
<td>224</td>
<td>1.7</td>
<td>1.7</td>
<td>219</td>
</tr>
<tr>
<td>KDEdoc</td>
<td>3</td>
<td>35k</td>
<td>31k</td>
<td>3</td>
</tr>
<tr>
<td>OpenOffice</td>
<td>42</td>
<td>.39</td>
<td>.37</td>
<td>38</td>
</tr>
<tr>
<td>OpenOffice</td>
<td>77</td>
<td>.42</td>
<td>.39</td>
<td>78</td>
</tr>
<tr>
<td>PHP</td>
<td>40</td>
<td>.23</td>
<td>.23</td>
<td>35</td>
</tr>
</tbody>
</table>

3 Benchmarks for Lab Tests

For the lab tests of MT components, texts have been selected for each domain and language pair. Their amount is such as to allow the definition of suitable tuning or testing sets. In the following paragraphs, we report statistics on the entire collections, as splits are left to cover specific experimental requirements.

3.1 Legal Domain

For the legal domain, benchmarks consist of documents randomly selected from the most recent release of the JRC-Acquis corpus. The following table shows the statistics for every language pair.
Table 3: statistics of Legal domain evaluation sets for lab tests.

<table>
<thead>
<tr>
<th>en-de</th>
<th>en-es</th>
<th>en-fr</th>
<th>en-it</th>
</tr>
</thead>
<tbody>
<tr>
<td>#sent</td>
<td>#sent</td>
<td>#sent</td>
<td>#sent</td>
</tr>
<tr>
<td>(k)</td>
<td>(k)</td>
<td>(k)</td>
<td>(k)</td>
</tr>
<tr>
<td>8.4</td>
<td>12.0</td>
<td>9.0</td>
<td>4.1</td>
</tr>
<tr>
<td>82.2</td>
<td>177.4</td>
<td>48.1</td>
<td>49.8</td>
</tr>
<tr>
<td>68.4</td>
<td>178.6</td>
<td>52.7</td>
<td>53.3</td>
</tr>
</tbody>
</table>

3.2 IT Domain

For the IT domain, Translated provided texts from real translation projects they conducted in the past. The following table shows the statistics for every language pair.

Table 4: statistics of IT domain evaluation sets for lab tests.

<table>
<thead>
<tr>
<th>en-de</th>
<th>en-es</th>
<th>en-fr</th>
<th>en-it</th>
</tr>
</thead>
<tbody>
<tr>
<td>#sent</td>
<td>#sent</td>
<td>#sent</td>
<td>#sent</td>
</tr>
<tr>
<td>(k)</td>
<td>(k)</td>
<td>(k)</td>
<td>(k)</td>
</tr>
<tr>
<td>9.8</td>
<td>9.8</td>
<td>9.8</td>
<td>9.8</td>
</tr>
<tr>
<td>229.3</td>
<td>232.3</td>
<td>233.8</td>
<td>233.1</td>
</tr>
<tr>
<td>210.6</td>
<td>262.7</td>
<td>250.1</td>
<td>240.0</td>
</tr>
</tbody>
</table>

4 Benchmarks for Field Tests

Four field tests will be carried out in two specific domains (Legal and IT) and for two language pairs (English to German and English to Italian). For each field test approximately 160,000 words will be translated broken down as follows:

<table>
<thead>
<tr>
<th>Domain</th>
<th>Translators</th>
<th>Words/translator</th>
<th>Total words</th>
</tr>
</thead>
<tbody>
<tr>
<td>EN&gt;DE</td>
<td>Legal</td>
<td>8</td>
<td>5,000</td>
</tr>
<tr>
<td>EN&gt;DE</td>
<td>IT</td>
<td>8</td>
<td>5,000</td>
</tr>
<tr>
<td>EN&gt;IT</td>
<td>Legal</td>
<td>8</td>
<td>5,000</td>
</tr>
<tr>
<td>EN&gt;IT</td>
<td>IT</td>
<td>8</td>
<td>5,000</td>
</tr>
<tr>
<td>Total</td>
<td>---</td>
<td>32</td>
<td>---</td>
</tr>
</tbody>
</table>
The large number of words that each translator will translate will limit the impact of the so-called 'warm-up period', i.e. the time necessary for the translator to get acquainted with the text and reach his/her standard translation speed. At the same time, the number of words will allow to complete the data collection for each field test in a reasonable amount of time (approx. 4 days).

One aspect to bear in mind is that each translator will use the MateCat tool differently and in general will translate following specific patterns and habits of working. In order to cope with the variability determined by each translator's behaviour and idiosyncrasy, a team of 8 translators per language pair and domain will be used, for a total of 32 translators.

4.1 Legal Domain

The text used for the field tests in the legal domain will be taken from recent documents released by the European institutions. The documents will be as homogeneous as possible in terms of length and syntactic and semantic complexity. Even though the translation of the selected documents is not publicly available online at the time of the field tests, it is indeed likely that multilingual portions of the documents are publicly available on the European Union websites. This may affect the results of the contrastive tests (see below) as the publicly available content is very likely to have been indexed and used for the training of commercial MT engines, such as Google Translate, hence improving their quality as compared to the MateCat MT engine.

4.2 IT Domain

For the IT domains, the text to translate will be extracted from software user manuals provided by clients of the commercial partner and not publicly available online nor elsewhere. As for the legal domain, the files will be as homogeneous as possible in terms of length and syntactic and semantic complexity.

5 Evaluation Methodologies

5.1 Lab Tests

In MateCat, domain adaptation and, to some extent, project adaptation of MT models is performed before the translators start working. This means that any upgrade of generic
models towards specific models can be measured by means of standard MT metrics, like BLEU and TER.

In particular, we will target automatic metrics that shown to correlate better with post-editing effort (O’Brien, 2011) and that allow measuring errors at the level of single segments. But MateCat involves MT engines that evolve over time as well, especially in the framework of user-adaptive MT. Evolving MT models pose the challenge of their effective evaluation. Since the ultimate goal of MateCat is to increase the productivity of human translators, the most accurate assessment methodology would be of course to run field tests. In this way, we could compare productivity of human translators receiving suggestions from an MT engine featuring dynamic adaptation against the productivity of human translators working with a static MT engine. Unfortunately, this evaluation is unfeasible during daily MT development, because too much expensive both in terms of cost and time. On the other hand, standard automatic MT metrics are unsuitable to track the dynamic behaviours. Metrics for measuring performance in the case of interactive MT, see for example (Khadivi, 2008), like Key-Stroke Ratio (KSR), Mouse-Action Ratio (MAR), Key-Stroke and Mouse-Action Ratio (KSMR) are expected to correlate with the productivity of human translators, but their computation requires the actual use of an interactive MT system, i.e. a field test.

In the SMART project¹, the evaluation of adaptive interactive MT is explored (Cesa-Bianchi et al., 2008). While no specific metric is proposed, the analysis is based on a plot of cumulative differences of BLEU score between a baseline and an adaptive system like in the following graph, where the behaviour of two adaptive systems, ITER1 and ITER2, are compared:

¹ http://www.smart-project.eu
These differences are computed sentence by sentence and present an interesting view of the dynamic change of an MT system. We have further elaborated on this idea in (Bertoldi et al., 2012), paper summarized in the following.

Other metrics like Character Error Rate (CER) and Translation Edit Rate (TER) would accurately predict the translators' productivity if the references were generated by using the CAT tool; on the contrary, references are usually generated from scratch based only on the source text and can thus be quite far from CAT-based translations, both lexically and syntactically. The Human-targeted variant of TER, HTER (Snover et al., 2006), needs human intervention, is therefore unfit to meet our requirements.

Recent work (S. O'Brien, 2011) on automatically predicting post-editing productivity has proposed the use of the General Text Matcher\(^2\) (GTM) metric in addition to TER. Roughly speaking, GTM computes an F-Score of maximal n-gram matches that rewards longer matches over shorter matches. This automatic metric will be investigated in MateCat, too.

In (Bertoldi et al., 2012) we proposed to use the percentage slope from the theory of learning curves to measure the learning ability of adaptive MT systems, as objective automatic

\(^2\) http://nlp.cs.nyu.edu/GTM
evaluation score for an MT system adapting over time. Before going on with the percentage slope, we need to define the experimental framework.

### 5.1.1 Dynamic Adaptation Framework

In the project scenario, the MT system, which is embedded in the CAT tool to increase the translators’ productivity, adapts over time by exploiting translations generated by the user. Each adapted instance is used to provide the user with translation suggestions for the next sentences. We refer to this process as dynamic (or incremental) adaptation to emphasize that adaptation happens continuously based on a stream of data.

From an abstract point of view, the framework of incremental adaptation can be summarized as follows:

1. before the process starts, an initial system is built on available data including a parallel corpus;
2. a stream of data becomes available (only the source is available yet) that is split into blocks of (not necessarily) similar size;
3. the first/next block of source text is considered;
4. the latest instance of the adapting system translates the source text of the current block;
5. the target part of the current block becomes available for use; note that in the CAT framework, the target part of a block is the translation post-edited by the user;
6. the system is adapted using the current parallel block and possibly all the previous ones;
7. the loop continues from step 3 until all blocks are processed.

In each adaptation step, all of the data available so far can be used, but no look ahead is possible. Note that, in principle, each block is translated with a different instance of the adapting system; hence, the same text occurring in two different blocks can be translated differently.

The performance of adaptive systems as sketched above is evaluated on different parts of the stream as opposed to the global evaluation used for static systems. We distinguish between two protocols that differ in their use of historic data. For block-wise evaluation only the translations of the most recent block are evaluated with respect to the correct translations once these become available. Any static automatic MT score, e.g. TER and BLEU, can be used, provided that it is reliable on a block of usually relatively small size. In contrast, in incremental evaluation the scores are computed on all blocks available so far. The translations of previous
blocks are kept fixed, i.e. blocks are not translated again once a newly adapted system becomes available, as this new system has already integrated this data into its models.

5.1.2 The Percentage Slope

Learning curves (see (Stump, 2002) for a detailed introduction) are mathematical models used to estimate efficiency gained when an activity is repeated. The learning effect was noted in industrial environment: the underlying notion is that when people repeat an activity, there tends to be a gain in efficiency. That is exactly the expected behaviour of our dynamically adapting MT system, thus we decided to exploit elements from the learning theory to measure the evolution of translation capability.

Several learning curve models have been proposed, but only two are in widespread use, the unit (U) model due to Crawford and the cumulative average (CA) model due to Wright (Stump, 2002). Both models are based on a common mathematical form:

\[ y = b \cdot a^x \]

where:
- \( a \) represents the theoretical labour hours required to build the first unit produced (a positive number);
- \( b \) represents the rate of learning (negative value, except for forgetting);
- \( x \) represents the number of an item in the production sequence (unit #1, #2, #3...).

The models differ in the interpretation of \( y \):
- \( U \): \( y \) is the labour hours required to build unit \( #x \);
- \( CA \): \( y \) is the average labour hours per unit required to build the first \( x \) units.

Since \( b \) is a mathematically appropriate but counter-intuitive number for describing the slope, the percentage slope \( S \) is typically used:

\[ S = 10^{b \log_{10}(2)+2} \]

\( S \) provides the rate of learning on a scale of 0 to 100, in percentage. A 100% slope represents no learning at all, while zero percentage reflects a theoretically infinite rate of learning. In practice, human operations hardly ever achieve a rate of learning faster than 70% as measured on this scale.
The correspondence between our block-wise evaluation with the U model, and the incremental evaluation with the CA model is straightforward. In the first case, $y$ is the number of errors done in the translation of the block $#x$; in the second case, $y$ is the average number of errors (that is the TER score or the 100-%BLEU score) made on the first $x$ blocks.

From a practical point of view, the sequence of scores can be provided while the adapting system is being used; the learning curve that best matches the sequence is then found\(^3\) and the percentage slope $S$ is computed.

### 5.1.3 Experimental assessment of the Percentage Slope

To assess its effectiveness, the percentage slope has been computed on errors committed by four MT systems that differently evolve over time. The following graph plots the TER score of each of them on the stream of input text to be translated that is also used to dynamically adapt the models: behaviours range from effective adaptation ($ada$ and $FGonly$ $fair$ $ada$ systems) to “forgetting” effect ($FGonly$ $unfair$ $ada$ system), passing through a not at all adaptation ($bsln$ $system$).

---

\(^3\) Notice that the best fitting learning curve can be estimated in the log scale with a simple linear regression analysis.
The $S$ values are collected in the following table.

```
<table>
<thead>
<tr>
<th>model</th>
<th>bsln</th>
<th>ada</th>
<th>FG-only ada</th>
<th>FAIR</th>
<th>UNFAIR</th>
</tr>
</thead>
<tbody>
<tr>
<td>U</td>
<td>100.4</td>
<td>96.9</td>
<td>96.2</td>
<td>107.2</td>
<td></td>
</tr>
<tr>
<td>CA</td>
<td>100.3</td>
<td>97.7</td>
<td>96.5</td>
<td>107.4</td>
<td></td>
</tr>
</tbody>
</table>
```

Expectations are confirmed:

- The system unable to learn ($bsln$) has an $S$ of 100%
- The two adapting system have an $S$ of 96-98%
- The system worsening its performance over time has an $S$ greater than 100%.

Therefore, we can state that $S$ exposes common behaviours of evolving SMT systems.

In order to give a hint for properly interpreting the values reported, we summarize the discussion in (Stump, 2002) about “typical learning slopes”. Operations that are fully
automated tend to have slopes of 100%, 70% if entirely manual, an intermediate value if mixed. In real industrial environments, the average slope depends on the type of manufacturing activity: for example, in aircraft industry it is about 85%, it ranges in 90-95% in electronics and in machining. Hence, a 96-98% slope as we measured in our experiments must be considered a significant learning ability of a fully automated system.

### 5.1.4 Open Issues

We have seen that in lab testing, in addition to the standard metrics commonly used by the MT community, we will compute the percentage slope to measure the learning ability of adaptive SMT systems. Anyway, some issues about lab evaluations still remain open. In particular, in MateCat, MT engines are used to compensate for the lack of generalization of the TM; hence, even in lab testing, in addition to stand-alone evaluations of MT, we should compare it with the TM. That is we should compare the usefulness of MT outputs against fuzzy matches of TMs.

It is worth noting that nowadays, in the framework of professional translators, payment of per word rates are linked to TM matching rates. This has caused commoditization of the actual translation work, and translators today are expected to track baskets of words and get paid at different rates in schemes such as the hypothetical one shown below:

<table>
<thead>
<tr>
<th>Matching Rate</th>
<th>Pay Rate per Word</th>
</tr>
</thead>
<tbody>
<tr>
<td>Repetitions</td>
<td>25%</td>
</tr>
<tr>
<td>100%</td>
<td>25%</td>
</tr>
<tr>
<td>85% to 99%</td>
<td>45%</td>
</tr>
<tr>
<td>75% to 84%</td>
<td>60%</td>
</tr>
<tr>
<td>0 to 74%</td>
<td>100%</td>
</tr>
</tbody>
</table>

We can wonder if, by using the MateCat tool which is equipped with both TM and MT, the actual post-editing effort is still closely tied to the matching rate or not. We will rely on crowdsourcing (Mechanical Turk) to evaluate the effort required to get either the MT or fuzzy matches from TM to a target quality level. This investigation will also provide useful hints on how to choose the suggestion to be displayed to users among those coming from the fuzzy matches and from the automatic translation.
5.2 Field Tests

The ultimate goal of the MateCat Tool is to increase the productivity of human translators in terms of words translated per hour, adaptability of the engine to specific domains, accuracy and consistency of TM matches and automatic suggestions, etc.

The evaluation-analysis cycles mentioned in the introduction will be critical throughout the duration of the project as they will inform the technical decision-making in two respects: (1) by measuring the impact of the individual components on the global performance and (2) by identifying the most important aspects to be improved from the human translator perspective.

Field tests will allow performing two different analysis of the MateCat tool. On the one hand, they will allow for the measurement of the productivity gains based on the collection and statistical analysis of translation data (productivity test). On the other hand, we will assess the usability of the MateCat tool by interviewing professional translators (usability study).
5.2.1 Productivity test

Productivity tests aim at measuring two key performance indicators:

1. Words per hour, i.e. the average translation drafting speed of translators.
2. Post-editing effort, which is the average percentage of word changes applied by the translators on the suggestions provided by the CAT tool.

In order to measure the productivity gains, four field tests will be carried out for two language pairs (EN>DE and EN>IT) and two domains (legal and IT). The field tests will be based on the process developed to complete the first field test which was conducted using a commercial CAT tool and MT/TM technology in order to define the baseline system and the processes that will be followed in further field tests to be carried out with MateCat.

For their very nature, field tests tend to present a high level of variability. Three main factors may affect the results of such tests:

- Characteristics of the text to translate (complexity, terminological density, length of sentences, formatting, etc.);
- Translators’ habit of working (experience with and usage of the translation software, ability to follow specific instructions, understanding of the job quality requirements, etc.);
- Translators’ knowledge of the text topic and terminology.

Accurately selecting the documents for the field tests and removing any significant differences in the text variability will reduce the impact of the first factor. By contrast, the impact of translators’ habits is more difficult to control as it depends on subjective features linked to the modus operandi of each translator.

To some extent, inter-user variability can be reduced by applying a number of constraints on how translators can work and by requiring translators to follow specific processes. However, experience shows that the translation process is highly subjective and that software constraints or detailed instructions are not enough to level out all differences in how translators work. In order to overcome such limitations, the field tests will be carried out employing the same translators for all tests and by selecting translators showing similar behaviour and translation productivity for detailed contrastive tests (see below).
D5.2 – Evaluation Plan

For all field tests, translators will be asked to follow specific instructions. They will have to translate the text assigned to them in a sequential order from one segment to the next, without skipping any segment in the process. Also, they will not be allowed to go back to a translated segment to edit it further and they will be required to deliver a draft translation (i.e. first editing of the target text without any review). These limitations are introduced to allow for the data collection to be as accurate as possible.

The data to be collected and analysed for each field test are:

- Source segments
- Matches from the TM or MT
- Target segments edited by the translator
- Time to edit each segment
- Percentage of words edited in each segment based on a comparison of the match provided by the system and the edited segment

The time to edit is computed by measuring the time difference from the moment when a request for MT and/or TM matches for a given segment is sent to the server to the moment when the human edited translation of the same segment is saved back on the server. The time difference is then used to compute the average words per hour translated by the translator.

The percentage of words edited in a segment is measured using a proprietary function developed by the commercial partner of the consortium. The function compares the match provided to the translator and the edited segment sent back by the translator and assigns a match percentage based on factors such as same words in the two segments and word order. Applying penalties based on factors such as formatting, tags, casing, etc then alters the match percentage. The function is designed to be as similar as possible to the matching algorithm used in commercial CAT tools.

All collected data will be stored in a database after the completion of the analysis stage. This will allow further analysis should a new statistical approach offer an alternative insight into the results of the field tests.

5.2.1.1 Statistical Analysis

Through field tests a large amount of observations about the users will be collected. Key performance indicators will be compared for each translator and each contrastive condition.
To assess the significance of the measured productivity gains, we will apply statistical tests to compute the probability they are obtained by chance. We will rely on so-called approximate randomized (or permutation) tests, which have recently shown to be more powerful than the classical t-tests and bootstrap methods (Riezler and Maxwell, 2005).

5.2.2 Contrastive test

While the overall success of the project will be measured by analysing any increase in translators’ productivity, tests with users in the loop and on real translation projects offer the opportunity to gain a better insight into how the development of the individual MateCat components affects the performance of the MateCat tool as a whole.

After each field test, a contrastive test will be carried out to analyse the performance of the MateCat tool with specific features enabled/disabled and, finally, to compare the MateCat MT server with Google Translate.

<table>
<thead>
<tr>
<th></th>
<th>Condition A</th>
<th>Condition B</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>First Contrastive</strong></td>
<td>Self-tuning MT enabled</td>
<td>Self-tuning MT disabled</td>
</tr>
<tr>
<td><strong>test</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Second Contrastive</strong></td>
<td>User-adaptive MT enabled</td>
<td>User-adaptive MT disabled</td>
</tr>
<tr>
<td><strong>test</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Third Contrastive</strong></td>
<td>Informative MT enabled</td>
<td>Informative MT disabled</td>
</tr>
<tr>
<td><strong>test</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Fourth Contrastive</strong></td>
<td>MateCat tool connected to MateCat MT server</td>
<td>MateCat tool connected to Google Translate</td>
</tr>
<tr>
<td><strong>test</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

As stated above, the results of any field tests can vary depending on a number of factors which can be controlled yet not completely removed and are mostly dependent on what we may call the human factor, i.e. the translators. In order to overcome such limitations, two translators per language displaying little variability in terms of productivity during the field tests will be used in the contrastive tests.
The contrastive tests will follow the same requirements as detailed here for the field tests, but will only involve two translators per test:

<table>
<thead>
<tr>
<th></th>
<th>Domain</th>
<th>Translators</th>
<th>Words/translator</th>
<th>Total words</th>
</tr>
</thead>
<tbody>
<tr>
<td>EN&gt;DE</td>
<td>Legal</td>
<td>2</td>
<td>5,000</td>
<td>10,000</td>
</tr>
<tr>
<td>EN&gt;DE</td>
<td>IT</td>
<td>2</td>
<td>5,000</td>
<td>10,000</td>
</tr>
<tr>
<td>EN&gt;IT</td>
<td>Legal</td>
<td>2</td>
<td>5,000</td>
<td>10,000</td>
</tr>
<tr>
<td>EN&gt;IT</td>
<td>IT</td>
<td>2</td>
<td>5,000</td>
<td>10,000</td>
</tr>
<tr>
<td>Total</td>
<td>---</td>
<td>8</td>
<td>---</td>
<td>40,000</td>
</tr>
</tbody>
</table>

5.2.3 Usability study

The MateCat project will introduce a number of changes to the usual translation workflow with current commercial CAT tools. Significant changes will affect both the information provided by the software to support the translation process (i.e. TM matches, MT suggestions, terminology aid etc.) and the user interaction with the software. The latter aspect will be subject to a usability study to determine the ease of use and efficiency of the MateCat tool.

The development of the MateCat tool will be accompanied by a continuous assessment of the user experience performed through:

- Informal interviews with staff of the commercial partner or part of the MateCat user group;
- Formal interviews with three teams of five professional translators on-site.
5.2.3.1 Informal interviews

The development of MateCat tool user interface will be subject to constant feedback from the personnel of the commercial partner from the onset. During the development process, account managers, project managers and professional translators will be asked to provide general feedback on a number of aspects such as look and feel, ease of use and general work flow. Interviews will allow to gather first impressions on the MateCat tool from professionals covering different roles in the translation industry and will inform the development process.

5.2.3.2 Formal interviews

Three teams of five professional translators working with the commercial partner who have never used the MateCat tool prior to the usability test will be selected to perform an in-depth study of the usability aspects of the MateCat tool. They will be interviewed at different stages of the development process on-site, at the commercial partner’s premises.

The interviews will be scheduled so as to precede each of the three releases of the MateCat tool. The focus of such interviews is to determine the usability of the MateCat Web-based tool and the effectiveness of its design from a professional user perspective. Following Nielsen (2003) definition of usability, the interviews will be focused on collecting information on the following aspects:

- **Learnability**, defining how easy it is for users to accomplish a list of tasks the first time they use the software;
- **Efficiency**, measuring how quickly users perform a list of tasks once they have learned how to use the tool;
- **Errors**, collecting data on the number and type of errors made by the users while performing the tasks and how they recover from the errors;
- **Satisfaction**, that is emotional response to the look and feel of the MateCat tool.

The list of tasks to be performed will cover all functionalities provided by the MateCat tool at the different stages of development. Given that each new version of the software will include new or updated features, the tasks to be accomplished will vary according to the progress in the MateCat tool development. Each translator will take part in only one test and will be asked to perform the same tasks as the other translators in his/her team.
What follows is a tentative list of tasks that translators will be asked to perform. The list will be updated to match the evolving features of the MateCat tool and, where necessary, to carry out a more in-depth analysis of tasks which have proven to be more complex to perform.

- Access the document to translate
- Identify the text to translate
- Open segment
- Check the TM/MT matches automatically provided
- Check the terminology matches automatically provided
- Translate segment
- Edit translated segment
- Approve translation
- Identify segment status
- Look up sub-segment matches (concordance)
- Look up terms
- Add terms to glossary
- Interact with other users
- Add comments to a segment
- Find comments added by other users
- Manage tags in source and target segment
- Export translated file
- Export translation memory file
- Export glossary
- etc.
Appendix 1. Draft of usability questionnaire

*Required Field

**Project Number** *

Please, enter the project number as indicated in your PO (e.g. 9691230-10081835)

**Browser**
Which browser did you use?
- Chrome
- Firefox
- Internet Explorer
- Other: [ ]

**UI - Segments**
Could you easily open/close the segment to translate?
- Yes
- No

**Readability**
Is the text to translate easy to read?
- Yes
- No

**UI - Shortcuts**
Were the keyboard shortcuts useful?
- Yes
- No
- Did not use them

**Edit area**
Can you easily identify the area where the translation is performed
- Yes
- No

**UI - Responsiveness**
Would you consider the MateCat tool to be...
- Very fast
- Fast
- Just OK
- Slow
- Very slow

---

Machine Translation Enhanced Computer Assisted Translation

D5.2 – Evaluation Plan

6 Appendix 1. Draft of usability questionnaire
D5.2 – Evaluation Plan

UI - Matches
Were the TM/MT matches presented clearly?
- Yes
- No
- No TM/MT matches received

Match Quality
Were the TM/MT matches helpful?
- Yes
- No

MT Suggestions
Were MT suggestions helpful?
- Yes
- No
- No MT matches received

Current Features
Does MateCat offer all the commands needed to perform a translation? If no, please specify in the comment section below.
- Yes
- No

Missing Features
What is the most important feature that MateCat currently lacks?
- Preview
- Tag Management
- Spellchecker
- Search & Replace
- Concordance
- Undo/Redo
- Segments Split & Merge
- Other: 

Comments
Please, provide any comments on what you feel could be improved in MateCat.

Submit
7 Bibliography

Statistical computer-assisted translation.
Ph.D. thesis, RWTH Aachen University, Aachen, Germany.
Advisors: Hermann Ney and Enrique Vidal.

Online learning algorithms for computer-assisted translation.
Deliverable 4.2, SMART project (FP6).
http://www.smart-project.eu/files/D42.pdf


Evaluating the learning curve of domain adaptive statistical machine translation systems.
In Proc. of WMT, Montreal, Quebec, Canada.

On some pitfalls in automatic evaluation and significance testing for MT.
In Proc. of ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Text Summarization, Ann Arbor, US-MI.

A study of translation edit rate with targeted human annotation.

All about learning curves.

Usability 101: Introduction to Usability.