This document is part of the Project "Machine Translation Enhanced Computer Assisted Translation (MateCat)", funded by the 7th Framework Programme of the European Commission through grant agreement no.: 287688.

Machine Translation Enhanced
Computer Assisted Translation

D5.4 – Second Report on Lab and Field Test

Authors: Nicola Bertoldi, Mauro Cettolo, Matteo Negri, Marco Turchi, Marcello Federico, Holger Schwenk

Dissemination Level: Public

Date: 15 November 2013

Grant agreement no.	287688
Project acronym	MateCat
Project full title	Machine Translation Enhanced Computer Assisted Translation
Funding scheme	Collaborative project
Coordinator	Marcello Federico (FBK)
Start date, duration	November 1st 2011, 36 months
Dissemination level	Public
Contractual date of delivery	October 31st, 2013
Actual date of delivery	November 15th, 2013
Deliverable number	5.4
Deliverable title	Second report on Lab and field tests
Type	Report
Status and version	Final, V1.1
Number of pages	28
Contributing partners	FBK, LEMANS
WP leader	Translated
Task leader	FBK
Authors	Nicola Bertoldi, Mauro Cettolo, Matteo Negri, Marco Turchi, Marcello Federico, Holger Schwenk
Reviewer	Christian Buck
EC project officer	Aleksandra Wesolowska
The partners in MateCat are:	Fondazione Bruno Kessler (FBK), Italy Université Le Mans (LE MANS), France The University of Edinburgh (UEDIN) Translated S.r.l. (TRANSLATED)

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

FBK
MateCat
Marcello Federico federico@fbk.eu
Povo - Via Sommarive 18 Phone: +39 0461 314 521
I-38123 Trento, Italy Fax: +39 0461 314 591

Copies of reports and other material can also be accessed via http://www.matecat.com

Executive Summary

This document reports on the lab and field tests carried out in the second year of the MateCat project. The field test aimed at evaluating the impact of self-tuning, user-adaptive and informative machine translation components on the user productivity. The field test employed the second version of the MateCat Tool developed by the industrial partner and MT engines developed by the research partners. The field test was carried out with professional translators, under very realistic working conditions, on two translation directions, English to Italian and English to French, and two linguistic domains: legal and information technology. The lab tests were instead run on all four translation directions covered by this project, which also include English to German and English to Spanish. The lab tests provide a detailed report on the performance of the single MT components that have been evaluated by the field test.

Contents

1 Introduction

Field tests are the most important evaluation activity of the MateCat project. They are carried with annual cadence and their main goal is to measure the impact of enhanced MT technology on the productivity of professional translators working with the MateCat Tool. Lab tests are carried out in parallel to field tests, too. Lab tests aim at measuring the performance of single components of the MT engines in isolation and by means of conventional automatic metrics. Besides allowing us to inspect the behaviour of MT components in a deeper way, we also run lab tests to verify that our methods generalise to other translation directions. The MateCat project is covering four translation directions, namely English to French, German, Italian, and Spanish, and two linguistic domains, namely information technology technical documentation and legal documents issued by the European Commission.

Since the start of the project, a preliminary field test was carried out at month 6, by employing a commercial CAT tool (SDL Trados) and a commercial MT engine (Google Translate). The goal of this test was to set a good reference baseline for the new MateCat Tool, especially in terms of achievable user productivity through the introduction of MT suggestions. The actual first field test was carried out later in October 2012 (month 12), with the aim of evaluating the impact of self-tuning MT on user productivity. As foreseen by the project work plan, the field test was limited to two translation directions, English-German and English-Italian. Deliverable 5.3 describes in detail the set-up of the first field test and reports on results. The outcome of the first field test suggested the consortium to reconsider the inclusion of the English-to-German translation direction in the forthcoming field tests. Indeed, although the partners are capable of developing close to state-of-the-art MT systems, the level of performance achieved for this direction does not yet seem adequate to allow for productivity gains through post-editing. During early Summer 2013, intermediate tests with professional translators were run with an improved English-German MT system and with an English-French MT system. The outcome of these experiments was the decision to run the future field tests with Italian and French, and to work with German and Spanish for the lab tests.

The second field test was run in October 2013 (month 24) and was organized in two consecutive sessions, lasting several days, in which translators had to work on different portions of a document. During the first session, called *warm-up session*, each translator had to translate a portion of the test document by post-editing suggestions generated by a domain-adapted (reference) MT engine. The warm-up session has different purposes: to let the translator get familiar with the MateCat tool, to evaluate the quality estimation component and, last but not least, to collect translations by the user in order to perform project adaptation. The warm-up session was completed by most translators in one working day. Then, the actual *field-test session* took place, during which the remaining part of the test document was translated. During the field-test session, which lasted three to four days, each translator post-edited MT suggestions coming from either from the reference MT engine or from an enhanced engine featuring self-tuning and user-adapted MT. In particular, off-line adaptation was performed by exploiting the translations collected during the warm-up session, while on-line adaptation directly took advantage of the post-edits produced during the field-test session.

At the end of the field test, the impact of these enhancements was measured by comparing productivity of translators on the document segments for which suggestions were respectively generated by the reference engine or by the enhanced MT engine. Productivity was measured by two key performance indicators: number of words translated per hour (time to edit) and human translation error rate (post-edit effort).

This field test also evaluated the impact of informative MT in terms of usability. In particular, a MT quality estimation system was integrated in the MT engine, that estimates a quality score for each MT suggestion. The user interface of the CAT tool was adapted to provide the quality score similarly to how the quality of the translation memory matches is displayed. As this feature does not directly impact on the quality of MT suggestions, it becomes difficult to measure its actual contribution to the user productivity. Hence, the impact of the quality estimation module was evaluated through a usability questionnaire.

Besides the second field test, this document reports also on the lab tests performed to evaluate the quality of the employed MT engines in isolation. Experiments on the translation directions of the field test, English-Italian and English-French, here named *primary directions*, were run on both domains with the same MT systems and test data of the field test. For the other two translation directions, English-Spanish and English-German, here named *additional directions*, experiments were run on the legal domain again, also on the same test document employed for the field test. As a main difference with the field tests, lab tests rely on already existing human translations that are independent from the MT system to be evaluated. Lab tests manly evaluated algorithms for self-tuning and user-adaptive MT. The impact of these algorithms is measured in terms of three automatic MT metrics: BLEU [Papineni et al., 2002], TER [Snover et al., 2006], and GTM [Turian et al., 2003].

In the next section we report on the field test carried out on the legal and IT domain with two translation directions: English-Italian and English-French. Then, we describe the lab tests carried out on the legal domain with all four translation directions covered by the project. This report ends with some discussion and concluding remarks in which the achieved results are compared against the goals set in the work plan and the definition of requirements for future developments of the MateCat project.

2 Field Test

2.1 Evaluation Data

For the Information Technology (IT) domain data were supplied by Translated. An already executed translation project from English was selected for which translations into Italian and French were available. As translations in the two languages were carried out with different CAT tools, some manual pre-processing was necessary to uniform the text segmentations of the documents across the two translation directions. Moreover, some text cleaning was performed to remove formatting tags and software code excerpts, which are not relevant for our field test. Finally, a single source document of about 1,956 segments and 17,800 source words was created, and split into two portions: one for the warm-up sessions (342 segments), and one for the actual field-test session (1614 segments).

For the legal domain a document was taken from the EC website, for which translations in the four languages covered by MateCat are already available. The document was also pre-processed so that the segments of the four versions were all aligned. The full document consists of 605 segments and 13,900 words, and was split into two portions: one for the warm-up session (133 segments) and one for the actual field-test session (472 segments).

Table 1 provides some statistics of texts to be translated during the Warm-Up (WU) sessions and the Field-Test (FT) session. Figures on target sides refer to human references. Note that for each domain, the document to translate is shared among all language-pairs. The small difference between WU Legal texts is due to a single segment not post-edited by all translators which had to be discarded.

direction	domain	test set	segments	tokens	
				source	target
en→it	IT	WU	342	3435	3583
		FT	1614	14388	14837
	Legal	WU	133	3082	3346
		FT	472	10822	11508
en→fr	IT	WU	342	3435	3902
		FT	1614	14388	15860
	Legal	WU	134	3084	3695
		FT	472	10822	12810

Table 1: Overall statistics on parallel data used for evaluation purposes: number of segments and running words of source and target sides. WU means *warm-up*, and FT stands for *field test*

2.1.1 Training Data

For training purposes, the linguistic resources listed in deliverable D1.1 have been used. Table 2 provides detailed statistics on the actual bitexts used for the estimation of models. The `train` entries refer to the whole in-domain texts, while the `FG` (foreground) entries refer to the subset of `train` data that were selected for the sake of project adaptation to the specific document to translate. The `trainOD` entry of the IT en→fr task refers to data selected from out-of-domain texts (Giga English-French, United Nation, and Common Crawl corpora) by using the in-domain text as seed; this was done to augment the amount of training data, since the size of in-domain text available for that language pair (15.4/17.9 million words) is about four times smaller than for the other tasks.

direction	domain	corpus	segments	tokens (M)	
				source	target
en→it	IT	train	5.4M	57.2	59.9
		FG	360k	3.8	4.0
	Legal	train	2.7M	61.4	63.2
		FG	180k	5.4	5.4
en→fr	IT	train	1.1M	15.4	17.9
		trainOD	1.2M	20.0	22.2
		FG	530k	8.6	9.5
	Legal	train	2.8M	65.7	71.1
		FG	180k	5.5	5.8

Table 2: Overall statistics on parallel data used for training purposes: number of segments and running words of source and target sides. Symbols k and M stand for 10^3 and 10^6, respectively.

The evaluation of MT engines in lab experiments was performed on the actual documents translated during the field test.

2.1.2 MT Systems

The SMT systems have been built upon the open-source MT toolkit Moses [Koehn et al., 2007]. The translation and the lexicalized reordering models are trained on the available parallel training data (Table 2); 5-gram LMs smoothed through the improved Kneser-Ney technique [Chen and Goodman, 1999] are estimated on the target side by means of the IRSTLM toolkit [Federico et al., 2008]. The weights of the log-linear interpolation model are optimized through the standard MERT procedure provided within the Moses toolkit.

Three different SMT engines have been tested; two of them, the baseline and that equipped with all available technologies, were actually used for generating suggestions during the field test. The third system has been tested to allow the factorization of the contributions given by the self-tuning MT and the user-adaptive MT. Here the list of model acronyms, with the corresponding meaning, used in the rest of the paper. Note that with *data selection*, *back-off* and *LM mixture* we refer to the methods described in [Cettolo et al., 2013a] and [Cettolo et al., 2013b] to perform the adaptation of the models to a specific project, while with *online* we refer to the cache-based online adaptation method presented in [Bertoldi et al., 2013].

`BG`: background model, trained on the whole domain specific training data

`FGtgt`: model trained on (parallel) data selected from the training data using the target side of the document translated during the WU session as seed corpus

`WU+FGtgt`: model trained on the concatenation of post-edits of the WU and of text used for `FGtgt`

`mix()`: mixture of LMs (linear interpolation)

`backoff()`: back-off of TMs

`online()`: online adaptation of TMs/LM

The three tested SMT engines are built on such models as follows:

`DA.STA`: the domain-adapted static system, based on `BG` models, used as reference baseline system

`PA.STA`: the static project-adapted system, i.e. with models `mix/backoff(WU+FGtgt,BG)`

`PA.DYN`: the adapted SMT engine where models are `online(mix/backoff(WU+FGtgt,BG))`, i.e. dynamically adapted

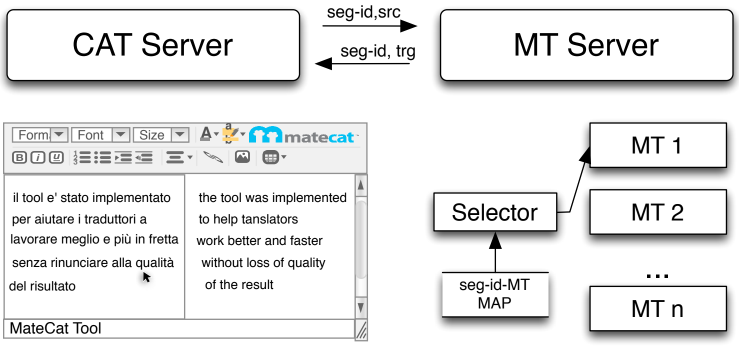
2.2 Field test protocol

The field test was conducted with 16 professional translators, 4 for each domain and translation direction. Each user works with the MateCat tool equipped with a private TM (initially empty) and one or more MT engines. The work of each translator (user) is organized in two sessions. During the first session (warm-up session), the user translates a portion of the document by receiving suggestions from the reference baseline MT system (`DA.STA`). During the warm-up the user also receives quality estimation scores by the MT system. After the warm-up phase, in the test phase the user translates the rest of the document by receiving and post-editing suggestions generated by two MT engines: the static domain-adapted MT system (`DA.STA`) and a dynamic project-adapted system (PA.DYN). In particular, for each segment only one suggestion is shown which is randomly chosen from the two engines. The translator is unaware not only of the source of suggestions but even that multiple engines are working. The mapping of segments to different MT engines is prepared in advance and is based on a quasi-random partition of the test segments in two sets, one for each MT engine. In fact, we continue exploring random partitions until we find one that passes a series of quality checks:

- the TER scores of the baseline system on the two sets must be similar
- the average length of the segments of the two sets must be similar
- the length distribution of the segments of the two sets must be similar

The above conditions are statistically checked. Random partitions over the test segments are also generated in a way that for each segment we can collected two post-edits of the DA.STA system and two of the PA.DYN system. Hence, we generate two random partitions using the above properties to make the MT assignments to the first two post-editors. Finally, for the remaining two post-editors we generate the corresponding complementary assignments. In other words, if for the first post-editor a given segment is translated with DA.STA, then for the third post-editor the same segment will be translated with PA.DYN, and so on.

The field test protocol is outlined in Figure 1.



il tool e' stato implementato per aiutare i traduttori a lavorare meglio e più in fretta senza rinunciare alla qualità del risultato

the tool was implemented to help tanslators work better and faster without loss of quality of the result

Figure 1: Implementation of the experimental protocol in the CAT tool. Single translation requests from the CAT server to the MT server are handled by different MT engines, according to a beforehand prepared table mapping segment codes to MT engines.

2.3 Evaluation criteria

For the field test we applied the same criteria used for the previous field test. Once a translator completes the assigned task, we download and process the log file generated by the MateCat tool. In particular, we split the time and post-edit statistics of all segments in two parts according to the source of the suggestion, that is system DA.STA or system PA.DYN.[1] For each collection of segments we compute and compare the average translation speed (words/hour) and the post-edit effort (HTER score). Translation speed is computed after removing possible noisy observations. In particular, the average speed of the union of all segments and edit times is computed for each condition after removing the 5% fastest and the 5% slowest post-edits. Finally, the HTER scores for each type of suggestion are computed as a weighted mean of TER scores by taking into account the length of each segment. Statistical significance of the differences in post-edit speed and effort is assessed via *approximate randomization* [Noreen, 1989], a statistical test well-established in the NLP community [Chinchor et al., 1993] and that, for the purpose of MT evaluation, has been shown [Riezler and Maxwell, 2005] to be less prone to type-I errors than the boostrap method [Efron and Tibshirani, 1993].

2.4 Troubleshooting

The execution of the field test was more difficult than expected. The same architecture, the CAT tool itself and the supporting MT engines, was already tested during the summer by running translation sessions with students. This positive experience was probably the main cause of an excessive confidence we had in the robustness of the system architecture, which is indeed rather complex. In fact, for each translator involved in the field test, the CAT tool has to be connected to a private TM and a private MT server. After the warm-up session, project adaptation for each translator is performed. Hence, for the field test, two MT systems have to be set-up behind the private MT server: one static domain-adapted system, which is in common for all the translators, and a dynamic project-adapted MT system, which is specific for the user. Hence, 4 domain-adapted MT engines have to be developed, to cover two translation domains and two translation directions, and 16 independent project-adapted MT systems, to cope with 16 translation tasks, for a total of 20 distinct MT engines to adapt and tune.

During the execution of the field test several problems occurred, the reason of some of which is still under investigation. In particular, during the tests with the legal domain, two major crashes of the MT servers caused the loss of the work of two of the eight translators, one for French and one for English. In fact, the translators completed their jobs but basically without post-editing MT suggestions. Other crashes occurred during the information technology test, due to an incorrect management of "<" and ">" characters[2] and to a bug in one of the core components of the on-line learning modules. Both issues were promptly spotted and fixed.

[1] Notice that segments for which the best suggestion came from the translation memory are discarded.

[2] The IT document of the field test was rather rich of programming language expressions, see excerpts in the Appendix.

In general, these issues caused losses of segments, that is segments for which translations are available but not useful for the field tests comparisons. In fact, after the field tests were completed, a careful analysis on both the log files produced by the MT server and by the MateCat tool was performed in order to isolate and discard unreliable segments which could not be used for our analysis. Finally, despite the fact that work of two translators could not be used, we could carry out the analysis of the results on a sufficient amount of segments for the remaining translators.

2.5 Results

Results of the field tests are reported in Tables 3-4.

	User	Time to edit (word/hour)				Post-editing effort (HTER)			
	ID	DA.STA	PA.DYN	p-value	Δ	DA.STA	PA.DYN	p-value	Δ
En-Fr	8045	144	150	0.376	4.02%	34.61	32.39	0.088	6.39%
	8047	102	121	0.045	15.32%	25.77	25.77	0.50	0.01%
	8048	79	74	0.208	-6.65%	45.51	40.97	0.009	9.98%
En-It	8021	195	221	0.167	11.59%	30.60	26.57	0.003	13.18%
	8022	147	149	0.455	1.50%	31.73	29.87	0.25	5.83%
	8024	65	97	0.004	32.79%	40.04	37.51	0.154	6.30%

Table 3: Time-to-edit and Post-editing effort for the field test on the Legal Domain. Measurements are taken on post-edits performed with the domain-adapted static MT system (DA.STA) and the project-adapted dynamic MT system (PA.DYN).

	User	Time-to-edit (word/hour)				Post-editing effort (HTER)			
	ID	DA.STA	PA.DYN	p-value	Δ	DA.STA	PA.DYN	p-value	Δ
En-Fr	7658	790	583	0.162	-26.13%	30.71	33.37	0.11	-7.97%
	7660	295	283	0.425	-3.83%	35.87	38.09	0.167	-5.84%
	7661	565	369	0.099	-34.69%	32.55	34.63	0.206	-6.02%
	7662	349	319	0.307	-8.49%	33.28	28.19	0.024	15.30%
En-It	7652	278	253	0.091	-9.02%	29.22	28.47	0.319	2.54%
	7655	147	149	0.455	1.50%	31.73	29.87	0.25	5.83%
	7656	674	511	0.269	-24.15%	34.99	36.60	0.157	-4.39%
	7657	418	560	0.155	25.36%	22.01	22.61	0.356	-2.66%

Table 4: Time-to-edit and Post-editing effort for the field test on the Information Technology. Measurements are taken on post-edits performed with the domain-adapted static MT system (DA.STA) and the project-adapted dynamic MT system (PA.DYN).

For each translator (user ID) we report time-to-edit and post-edit figures corresponding to the two sources of MT suggestion, the percentage of the differences between such figures, and the p-values of the hypothesis that the observed differences are random.

For the legal domain, the overall results for both translation directions are quite positive. All translators but one (User ID 8047) improved their productivity in terms of post-edit effort, thanks to project adaptation and on-line learning. Observed improvements range from 6% to 13% relative, and are statistically significant at level 0.01 in 3 cases out of 5. Concerning translation speed, 5 translators out of 6 seem to benefit from the improved MT engine. Remarkably, translation speed and post-edit effort are not highly correlated with each other. In fact, there seems to be a much higher variance inter and intra translators on this dimension, which reflects in the fact that only differences above 15% are statistically significant at level 0.05.

Results on the IT domain are more controversial. In absolute terms, translation speed figures look higher than for the Legal domain, which means that the task was less complex for the translator (probably less domain-specific terminology and complex sentences to translate).[3]. From the post-edit effort side, the HTER figures are in the range of those observed for the legal document. Quite surprisingly, concerning the impact of self-tuning and user adaptive MT, French translators not only did not benefit from these components, three out of four translators got worse HTER scores, and all of them lowered they translation speed. For the Italian direction, results look slightly better: the impact of MT adaptation on post-edit effort is moderate, with two out of four translators improving their HTER (though not significantly). On the side of translation speed, most of the observed variations are negative although not statistically significant.

These controversial results on the IT domain suggest that part of the project adaptation and on-line learning procedures went wrong. While a careful inspection of the log file showed that all the steps were applied correctly a more detailed analysis of the data used for adaptation was performed during the lab tests, which is reported in Section 3.

2.6 Quality Estimation

According to the MateCat agenda, the objectives for the second year of the project included the integration of a Quality Estimation (QE) component in the CAT tool and its evaluation by post editors operating in normal working conditions. Along this strand of activities, the target evaluation result (Milestone 4 in the MateCat Technical Annex) was "*40% of user acceptance of informative MT in field test*". The remainder of this section reports on how the two objectives have been timely accomplished.

[3] See excerpt of the document in the Appendix.

2.7 Integration of the QE component

As a result of the research activity on QE carried out within WP3 (see *D3.1 - First Report on Informative MT*), the first version of a component for the automatic assessment of translation quality has been deployed and integrated in the MateCat tool. Trained on a corpus of *[source, target, post-edited target]* tuples for a given language pair, the system is designed to return at run time a *QE score* for new, unseen *[source, target]* pairs.

The possible strategies to exploit the automatically computed predictions include: *i)* filtering out low-quality MT suggestions based on a fixed threshold, and *ii)* present all MT suggestions together with the computed QE scores. For the first integration of QE functionalities in MateCat we opted for the latter (more conservative) solution, aimed to help post-editors in selecting the suggestions (either from the MT or the TM) that are more suitable for manual correction. This decision is also motivated by the difficulty to set arbitrary quality thresholds as discussed in [Turchi et al., 2013].

To implement and evaluate such strategy, two QE components have been deployed for the translation directions addressed by the second year field test: English-Italian and English-French. The two components have been trained using the *[source, target, post-edited target]* tuples collected during the first field test (see *D5.3 – First report on lab and field tests*). To start with, only one domain (Information Technology) has been selected, leaving a comprehensive evaluation including Legal data as a task for the next round of tests. For both language directions, the collected data provide four post-editions for each automatic translation (282 target sentences for English-Italian and 353 fro English-French). However, since the design of effective strategies to capitalize on the output of multiple post-editors falls out of the scope of this preliminary integration effort, only one post edition for each target sentence has been considered.[4] To do this, the four post-editors have been ranked according to their average HTER (considered as an indicator of their post-editing behaviour). Then, removing the most aggressive (highest average HTER) and the most conservative post-editors (lowest HTER) training data were randomly collected from the two remaining post-editors. Finally, the collected data were used to train an SVR regressor (using an RBF kernel and the 17 “baseline” features extracted with the QuEst[5] tool for translation quality estimation [Specia et al., 2013]) following the modalities described in D3.1.

[4] In previous experiments we discovered that learning from all the available data is not trivial. Often, in fact, the high variability of post-editions of the same target sentence (due to the variability of subjective correction strategies) results in a noise the standard algorithms can not effectively handle. A recently proposed method to jointly learn from the output of multiple annotations [Cohn and Specia, 2013] will be considered as a future improvement direction.

[5] `http://www.quest.dcs.shef.ac.uk/`

2.8 Evaluation of the QE component

The two QE modules previously described have been *extrinsically* evaluated in the second MateCat field test.[6]

Aiming at a focused evaluation with limited impact on other system functionalities subject to evaluation, the presentation of QE scores has been activated only for the warm-up phase. In this phase, eight users (four for English-Italian and four for English-French) translated in one day 342 sentences for each translation direction. For each source sentence, both TM and MT were provided (the former with the standard fuzzy match score, the latter with the estimated QE score).

At the end of the warm-up, users were presented with a short web-based questionnaire aimed to collect their opinions about the potential usefulness of QE scores, and measure their satisfaction concerning the actual scores associated to the translated segments. The questionnaire contained the following three items (statements), on which users had to explicitly mark their agreement level with a 1-to-5 *Strongly_Disagree-Strongly_Agree* response scale:

1. **Opinion**. The estimation of the machine translation quality for each suggestion is useful in your work.
2. **Correlation**. For the segments that I translated, I reckon that the quality estimation reflects the actual quality of the MT suggestion (i.e. a high quality estimation value identifies a good suggestion from the MT)
3. **Usefulness**. The quality estimation score helps me to quickly identify whether a suggestion from the MT is useful to produce the translation

The results of the questionnaire are summarized in Figure 2. In particular, on the left side it shows the distribution of responses in the 1-to-5 scale while, on the right side, it shows a more coarse distribution into positive, neutral, and negative judgements. Overall, the figures demonstrate the general satisfaction of the users. While the agreement on the high potential of the QE functionality to simplify translators' work is very high (75% of the users expressed a positive judgement), the performance of our QE components seems still lagging, indicating a large room for improvement to fully meet such user demand. The positive judgements about the actual *correlation* and the *usefulness* of QE scores are in fact 50%, with three neutral and one negative opinion about the former and four neutral opinions about the latter. However, regarding the second objective related to QE research within MateCat ("*40% of user acceptance*

[6] In addition to that, D3.1 reports on positive *intrinsic* evaluation results achieved by the joint participation of two consortium members (FBK and the University of Edinburgh) in the shared QE task organized within the $8^{th} Workshop\ on\ Statistical\ Machine\ Translation.$

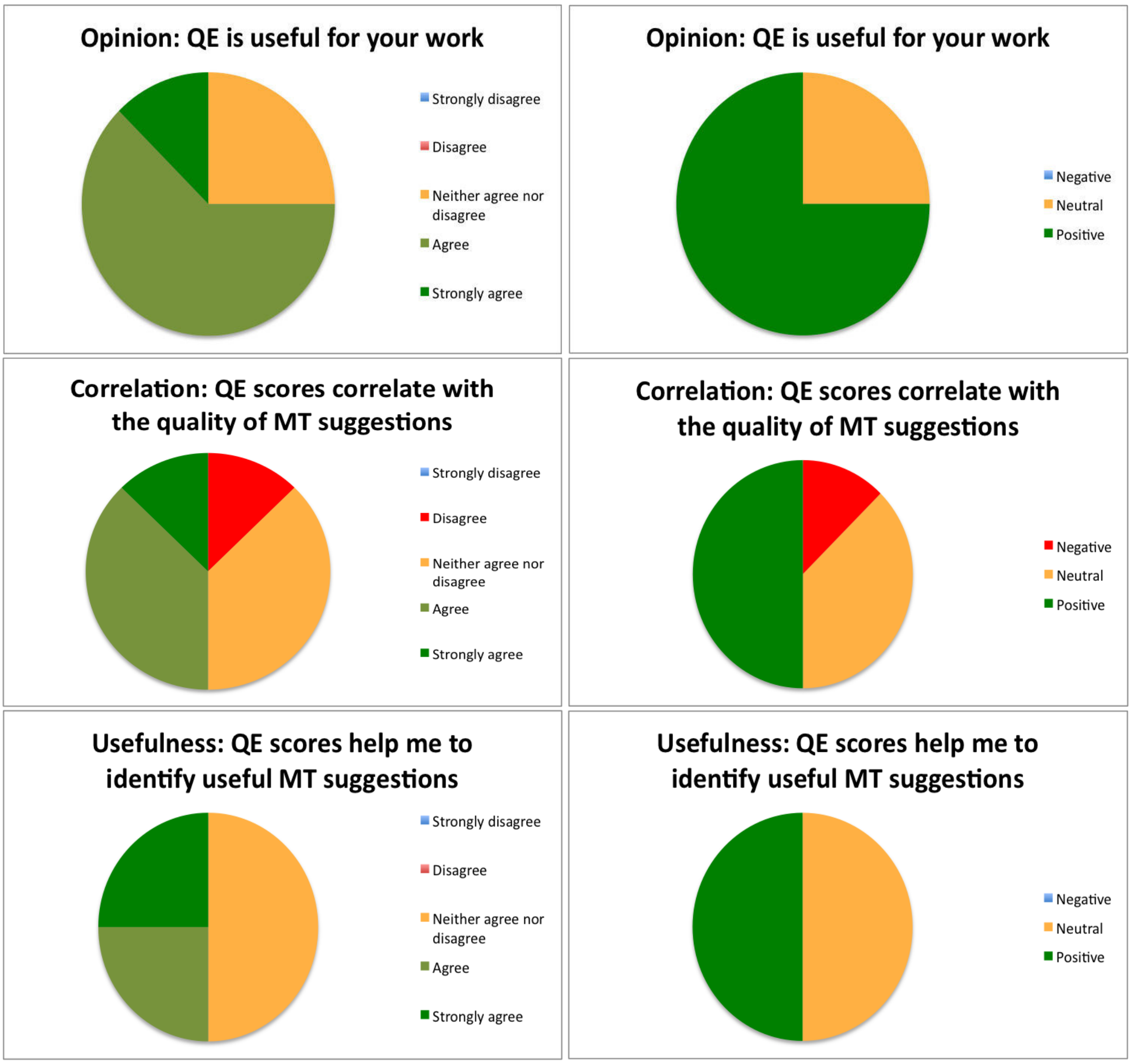


Figure 2: Fine (left column) and coarse (right column) opinion, correlation and usefulness user judgements about quality estimation of machine translation suggestions.

of informative MT in field test") the outcomes of this analysis show that the Year 2 achievements are in line with the project roadmap.

	English-Italian	English-French
PE1	22.39	26.38
PE2	21.34	25.33
PE3	23.48	25.53
PE4	22.66	26.09

Table 5: Mean Absolute Error (MAE) results calculated for each user (8 in total) working on two translation directions (respectively four translators for English-Italian and four for English-French).

Besides the positive and timely results in terms of user satisfaction, a quantitative analysis of the performance of our QE components brings additional interesting conclusions. On one side, it has to be remarked that strict performance results are worse compared to the state-of-the-art. As shown in Table 5, in fact, the Mean Absolute Error (MAE) calculated over the four post-editors working on English-Italian translations ranges from *21.34%* to *23.48%*, while for the English-French translations the MAE values range from *25.53%* to *26.38%*.[7] This is not surprising if we consider that the amount of training data available for our experiments which, for both language pairs, is around 10% of the data provided to WMT participants. Nevertheless, on the other side, it is interesting to note that such distance from the state-of-the-art does not significantly affect users' perception about the contribution of QE scores to speed-up and simplify their work. Considering the lack of studies on the real meaning of a given QE performance level (in terms of MAE), and its relation with the actual human quality standards and expectations, this is *per se* an interesting finding. In light of these considerations, and considering the large room for improvement to reduce the performance gap with respect to the top WMT systems, the possibility to raise the proportion of satisfied users above the current 50% is concrete and within the reach of the efforts planned for the third year of MateCat (*e.g.* the integration of the effective word alignment features we successfully explored in [de Souza et al., 2013a,b].

3 Lab Test

This section reports on lab experiments carried out for assessing the techniques implementing the self-tuning MT and the user-adaptive MT. Concerning the two language pairs of the field test - the primary directions - experiments involved both the IT and the Legal documents; for

[7] This year, for instance, the lowest MAE achieved on English-Spanish at the WMT QE shared task [Bojar et al., 2013] was 12.42%. Our joint FBK-UEdin submission to the same task achieved an MAE of 14.38

the additional directions, only the Legal domain has been considered. Lab tests were performed on exactly the same documents of the field test.

In the following subsections, we report results on the primary translation directions, then on the additional translation directions, and finally we report on a deeper analysis we carried out in order to explain the poor results on the information technology domain.

3.1 Primary Directions

Lab tests on the primary directions are based on the same training data, adaptation data, evaluation data and MT system setting employed for the field test. The only difference is that adaptation and evaluation scores rely on human reference translations collected independently from the MateCat tool.

3.1.1 Results

Table 6 collects BLEU, TER and GTM scores computed on the field test documents with respect to human references of the three systems for each of the four translation tasks.

pair	MT engine	IT domain			Legal domain		
		BLEU	TER	GTM	BLEU	TER	GTM
en→it	DA.STA	55.3	29.2	77.8	31.0	53.1	61.8
	PA.STA	55.4	28.7	78.2	35.0	49.1	64.6
	PA.DYN	59.7	24.6	80.9	36.5	48.1	65.6
en→fr	DA.STA	41.3	38.3	69.5	33.9	52.2	63.0
	PA.STA	41.4	37.9	69.9	36.4	49.1	65.1
	PA.DYN	41.9	38.0	70.5	39.5	46.4	67.6

Table 6: Overall performance of MT engines with respect to human references on the documents of the FT session.

For the Legal domain, the improvements over the baseline system yielded by the two adaptation techniques are impressive: the cumulative relative gain in terms of BLEU is about 17% for both the en→it and en→fr tasks; on the other metrics, it is lower but still around 10%. More specifically, the self-tuning MT is very effective for both language pairs, whereas the user-adaptation yields significant gains for en→fr, a bit smaller for en→it.

On the contrary, in the IT domain the only significant gain is observed with the dynamic adaptation component for the Italian direction. Remarkably, project adaptation does not impact for Italian nor for French (0.1 BLEU increase), and for French dynamic adaptation does also impact very marginally (0.5 BLEU increase). As these results are very different from what was observed with the experiments conducted so far with the previous field test data, a deeper analysis was conducted which is reported in Subsection 3.3.

3.2 Additional Directions

3.2.1 Data

As for the primary pairs, also for the additional pairs the linguistic resources listed in deliverable D1.1 have been used for training purposes. Table 7 provides statistics on actual bitexts used for model training in the Legal domain, which is the only domain considered for these language pairs.

direction	corpus	segments	tokens (M)	
			source	target
en→es	train	2.3M	56.1	62.0
	FG	180k	5.6	6.1
en→de	train	5.8M	140	131
	FG	1.9M	49.3	45.4

Table 7: Overall statistics on parallel data used for training purposes: number of segments and running words of source and target sides. Symbols k and M stand for 10^3 and 10^6, respectively.

Lab tests have been carried out using the same documents selected for the field test (Table 8). Also in this case the minor difference between the WU documents is due to few deletions caused by missing post-edits.

direction	test set	segments	tokens	
			source	target
en→es	WU	131	3007	3574
	FT	472	10822	12699
en→de	WU	133	3082	3125
	FT	472	10822	10963

Table 8: Overall statistics on parallel data used for evaluation purposes: number of segments and running words of source and target sides.

3.2.2 MT Systems

Concerning the SMT engine for the en→es pair, it is similar in all respects to those developed for primary directions.

Several efforts were made to improve the translation quality from English into German. First, a syntax system was build, based on Moses. Although the BLEU scores were lower than those of the phrase-based systems, the grammatical structure and coherence of the translations seemed to be slightly better. This could lead to a lower post-editing effort. After an analysis of the advantages and drawbacks of an English/German syntax system, we decided not to continue with it since its integration into the MateCat workflow is much more complicated than for a phrase-based system, e.g. performing incremental updates or calculating comparable confidence measures. Second, we tried to improve the English/German phrase-based system by pre-processing the English sentences. The idea was to reorder the English sentences in a way that the word order is closer to the one of the German translation. By these means, we hope to “simplify” the task of the phrase-based decoder, in particular with respect to long distance word movements. The English and German sentences were parsed and syntax trees created. A linguist analysed these trees and formulated a set of reordering rules which were applied to the training and test data. Unfortunately, these experiments did not result in significant improvements either. Finally, we identified a new development set which is more similar to the data used during the field test. The results reported below are based on this system.

3.2.3 Results

Table 9 reports automatic scores measured for the two additional language pairs in the Legal domain.

pair	MT engine	Legal domain		
		BLEU	TER	GTM
en→es	DA.STA	35.5	50.7	65.7
	PA.STA	36.4	50.2	65.6
	PA.DYN	39.5	46.9	68.2
en→de	DA.DYN	19.3	65.0	52.6
	PA.STA	20.1	64.7	52.8
	PA.DYN	21.4	62.5	55.1

Table 9: Overall performance of MT engines with respect to human references of D1.

Concerning the en→es direction, the cumulative improvement is similar to that observed for the primary language pairs in the Legal domain; in this case, the user adaptive MT is definitely more effective than the self-tuning MT.

Not surprisingly, the automatic metrics are much lower when translating into German, about 20 BLEU points in comparison to more than 30 for Italian, French and Spanish. Nevertheless, we can observe that the adaptation techniques are performing correctly. We observe an improvement in the BLEU score of 0.8 points when project adaptation is performed despite the fact that the new development set is already very similar to the field test data. Performing incremental adaptation gives an additional improvement of 1.3 BLEU points.

3.3 Analysis and Discussion

Lab test results for the IT domain were not as expected, especially for the en→fr direction. In this section, we investigate possible reasons for this outcome, by focusing on the most problematic language pair.

Self-tuning MT - Adaptation on the test project is done on the warm-up (WU) document with the aim of adapting the in-domain models towards the specific text to translate. The underlying assumption is that the text translated during the WU period well represents what will be translated in the field-test (FT) session. Table 10 provides some figures that seems to contradict this assumption.

test	RR	PP/OOV%		
set		DA.LM	PA.LM	PA_{FT}.LM
WU	29.5	242/4.2	-	-
FT	23.9	441/5.5	428/5.3	404/5.3

Table 10: Repetition rate (RR) of warm-up (WU) and field-test (FT) documents and their perplexity (PP) and out-of-vocabulary (OOV) word percentage, computed with language models of the domain-adapted (DA) system, and of two variants of the project adapted (PA) systems.

The column RR reports the repetition rate [Bertoldi et al., 2013], a measure of the repetitiveness of a text, of WU and FT documents on the target side (French): a 20% relative difference is a first warning. The differences in terms of perplexity (PP) and out-of-vocabulary (OOV) word percentage are even larger; the column DA.LM shows the two figures of the same documents computed on the language model (LM) of the domain adapted engine; in this case, the "difficulty" is greater for the FT document than for the WU document by almost 50% in terms of PP, and more than 20% in terms of OOV percentage.

The variant of the project adaptation adopted for preparing the models for this field test consists in selecting the most representative fraction of the whole training data by using the target side of the WU document as seed. Previous experiments [Cettolo et al., 2013b] showed that such a seeding is pretty equivalent to use the union of the source side of both the WU and FT documents. This holds under the assumption of a similarity of WU and FT texts. The columns PA.LM and PA_{FT}.LM give the PP/OOV values of the FT document on the LMs estimated by using the data selected in the two above mentioned ways, respectively. First of all, it results that the LM employed in the PA engine of the field test improves only marginally the LM of the domain adapted baseline (PP improves from 441 to 428, 3% relative). Secondly, the gain yielded by exploiting also the source side of the FT document for the data selection, although quite small (PP improves from 428 to 404, 6% relative), is definitely larger than in previous experiments reported in [Cettolo et al., 2013b]. These two outcomes undoubtedly show that the warm-up document does not represent well the FT text.

User-adaptive MT - Online adaptation with the `DA.DYN` system was performed by using the cache-based language and translation models [Bertoldi et al., 2013]. These models require one feature weight each that are tuned with the Simplex algorithm [Press et al., 1988] which finds weights by maximizing the translation score of the MT engine on a development set.

The development set employed for tuning the cache model weights was the warm-up document (IT domain, en→fr task) of the 2012 field test. This choice allowed us to run the weight estimation in advance, even before the 2013 field test. Moreover, we had evidence that any reasonable cache model weights were quite effective, provided their values are compatible with those weighting the other models in the log-linear interpolation defining the SMT model. Hence, we run weight tuning just for making all weights consistent with each other. Unfortunately, we discovered that this was a poor choice. Table 11 reports MT scores on the FT document by the PA.DYN engine employing cache model weights estimated on three different development sets.

dev set	BLEU	TER	GTM
WU 2012	41.9	38.0	70.5
WU 2013	48.4	34.9	74.5
FT 2013	51.3	32.1	75.9

Table 11: Overall performance of PA.DYN MT engine on field-test (FT) document by using weights of cache-based models estimated on different development sets.

The scores in the first row are the same of the corresponding entry in Table 6. The second row provides the scores we would have obtained if we had run the tuning on the current warm-up document of this year. The last row shows the unfair and upper-bound scores when the field-test document itself is used for tuning.

Clearly, although warm-up text does not well represent the field-test document as showed in the previous paragraph, it allows to estimate more effective weights than the warm-up document of 2012. In fact, the warm-up text of this year consists of a collection of small documents very different from those forming the field test, but from the same IT company, while the warm-up text of 2012 is from another company. This could probably explain the disappointing result.

Discussion - The investigation reported above suggests that the reason for the lack of improvement of the adapted models over the domain-adapted baseline models is the mismatch between the adaptation/tuning text and the actual test document.

Concerning self-tuning MT, our implementation is quite conservative and in the worst case it does not change at all the baseline quality; of course, we have learned that in any case of possible mismatch between adaptation and test data, it is important to exploit as much as possible the actual (source part) of the test data.

Concerning the User-adaptive MT, we implemented an aggressive method and probably underestimated the importance of the weight tuning step. Our plan for the future is to make the procedure safer by dynamically adjusting the cache model weights on post-edits as soon as they became available.

4 Discussion and Conclusion

The second field test was run as planned by using the second version of the MateCat Tool and MT engines developed by the consortium. The overall preparation of the field test was rather smooth and efficient, and improvements introduced in the protocol on the basis of the experience of the first field test, in order to reduce secondary effects in the contrastive tests, have proven to be effective. In particular, running two different MT engines on two different portions of the same document, by randomly interleaving the engines along the segments of the same document, was proven in previous internal tests as a viable solution to run contrastive tests. Unfortunately, the execution of the field test was rather difficult due to hardware and network issues, a software bug that was timely fixed, and a particularly difficult translation project selected for the information technology domain. In particular, the split into adaptation and evaluation data applied to the project document resulted in very distant data sets, thus violating the core assumption of adaptation data. Hence, both the project adaptation procedure and the tuning of the on-line learning model resulted in models that are worse than the more neutral reference system. A deep analysis of the results was very useful to identify problems and to envisage more robust adaptation procedures.

According to the work plan, the goals for the second year were to achieve 10% MT quality and productivity improvement, with self-tuning and user-adaptive MT, and 40% of user acceptance of informative MT. Despite the above issues, most of the success criteria set were met for the legal domain. In particular, the field test has reported productivity improvements on the post-edit effort side, on two translation directions, by 5 translators out of 6, with HTER reductions ranging from 6% up to 13%. More significant are the improvements measured in terms of MT quality according to automatic metrics. Static and dynamic adaptation methods, on the legal domain, resulted in BLEU score relative improvements above 10% for all language pairs. As regards user acceptance of informative MT, the integration of quality estimation functionalities has been positively evaluated by means of a questionnaire that revealed a level of satisfaction above the target 40% set as a Year 2 milestone.

Future work will be in the direction to anticipate and distribute the next field test over a longer period. Moreover, to cope with the increased complexity of the MT systems and the complexity introduced by the field-test protocol, efforts are underway to elaborate a detailed check list of conditions and tests to verify before and during the execution of the test, in order to mitigate the chance and impact of possible technical issues. As a further improvement, the next evaluation rounds will address a tighter integration (and broader evaluation) of quality estimation functionalities. To this aim, the first challenging task will consist in deploying effective methods to learn from small amounts of data, possibly by taking advantage of users feedback in an incremental fashion.

References

Nicola Bertoldi, Mauro Cettolo, and Marcello Federico. Cache-based Online Adaptation for Machine Translation Enhanced Computer Assisted Translation. In *Proceedings of the MT Summit XIV*, pages 35–42, Nice, France, September 2013.

Ondrej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. Findings of the 2013 Workshop on Statistical Machine Translation. In *Eighth Workshop on Statistical Machine Translation*, WMT-2013, pages 1–44, Sofia, Bulgaria, 2013. URL `http://www.aclweb.org/anthology/W13-2201`.

Mauro Cettolo, Nicola Bertoldi, and Marcello Federico. Project adaptation for mt-enhanced computer assisted translation. In *Proceedings of the MT Summit XIV*, Nice, France, September 2013a.

Mauro Cettolo, Christophe Servan, Nicola Bertoldi, Marcello Federico, Loïc Barrault, and Holger Schwenk. Issues in incremental adaptation of statistical mt from human post-edits. In *Proceedings of the MT Summit XIV Workshop on Post-editing Technology and Practice (WPTP-2)*, Nice, France, September 2013b.

Stanley F. Chen and Joshua Goodman. An empirical study of smoothing techniques for language modeling. *Computer Speech and Language*, 4(13):359–393, 1999.

Nancy Chinchor, Lynette Hirschman, and David D. Lewis. Evaluating message understanding systems: An analysis of the third message understanding conference (muc-3). *Computational Linguistics*, 19(3):409–449, 1993.

Trevor Cohn and Lucia Specia. Modelling Annotator Bias with Multi-task Gaussian Processes: An Application to Machine Translation Quality Estimation. 2013.

José G. C. de Souza, Christian Buck, Marco Turchi, and Matteo Negri. FBK-UEdin Participation to the WMT13 Quality Estimation Shared-Task. In *Proceedings of the 8th Workshop on Statistical Machine Translation (WMT'13)*, Sofia, Bulgaria, 2013a.

José G. C. de Souza, Miquel Esplà-Gomis, Marco Turchi, and Matteo Negri. Exploiting Qualitative Information from Automatic Word Alignment for Cross-lingual NLP Tasks. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, 2013b.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall, 1993.

Marcello Federico, Nicola Bertoldi, and Mauro Cettolo. IRSTLM: an Open Source Toolkit for Handling Large Scale Language Models. In *Proceedings of Interspeech*, pages 1618–1621, Brisbane, Australia, 2008.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic, 2007. URL `http://aclweb.org/anthology-new/P/P07/P07-2045.pdf`.

Eric W. Noreen. *Computer Intensive Methods for Testing Hypotheses: An Introduction*. Wiley Interscience, 1989.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association of Computational Linguistics (ACL)*, pages 311–318, Philadelphia, PA, 2002. URL `http://aclweb.org/anthology-new/P/P02/P02-1040.pdf`.

W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling. *Numerical Recipes in C*. Cambridge University Press, New York, NY, 1988.

Stefan Riezler and John T. Maxwell. On some pitfalls in automatic evaluation and significance testing for MT. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 57–64, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W/W05/W05-0908`.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. A study of translation edit rate with targeted human annotation. In *5th Conference of the Association for Machine Translation in the Americas (AMTA)*, Boston, Massachusetts, August 2006.

Lucia Specia, Kashif Shah, Jose G.C. de Souza, and Trevor Cohn. QuEst - A Translation Quality Estimation Framework. In *51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, ACL-2013, pages 79–84, Sofia, Bulgaria, 2013. URL `http://www.aclweb.org/anthology/P13-4014`.

Marco Turchi, Matteo Negri, and Marcello Federico. Coping with the Subjectivity of Human Judgements in MT Quality Estimation. In *Proceedings of the 8th Workshop on Statistical Machine Translation (WMT'13)*, Sofia, Bulgaria, 2013.

Joseph P. Turian, I. Dan Melamed, and Luke Shen. Evaluation of machine translation and its evaluation. In *Proceedings of the MT Summit IX*, 2003.

Appendix A

Excerpt of the field test translation project: IT domain, English-Italian, source and reference

```
a] Problem statement date:
a] Enunciazione del problema della data:

To get a date which is 5 days before the system date.
Ottenere una data precedente di 5 giorni la data di sistema.

CALCULATEDATE(sysdate(), 'd', -5) will return '01/13/2013' \
when sysdate()= '01/18/2013'.
CALCULATEDATE(sysdate(), 'd', -5) restituisce '01/13/2013' \
quando sysdate()= '01/18/2013'.

b] Problem statement date:
b] Enunciazione del problema della data:

To get a date which is 4 weeks after system date .
Ottenere una data successiva di 4 settimane la data di sistema.

CALCULATEDATE(sysdate(), 'd', 4*7) Formula at offset 4*7=28 days
CALCULATEDATE(sysdate(), 'd', 4*7) Formula con scostamento 4*7=28 giorni

The offset can be a formula.
Lo scostamento può essere una formula.

In the example the result of offset value is 28.
Nell'esempio il risultato del valore di scostamento è 28.

c] Problem statement:
c] Enunciazione del problema:

Retirement date of an employee considering he retires at age 60 years.
Data di pensionamento di un dipendente considerando va in pensione all'et
```

Excerpt of the field test translation project: legal domain, English-Italian, source and reference

Exchange of classified information with third States and international organisations
Scambio di informazioni classificate con Stati terzi ed organizzazioni internazionali

Where the Council determines that there is a need to exchange EUCI with a third State or international organisation, an appropriate framework shall be put in place to that effect.
Qualora il Consiglio ravvisi la necessità di scambiare ICUE con uno Stato terzo o un'organizzazione internazionale, è posto in essere a tal fine un quadro appropriato.

In order to establish such a framework and define reciprocal rules on the protection of classified information exchanged:
Per stabilire tale quadro e definire regole reciproche sulla protezione delle informazioni classificate scambiate:

the Union shall conclude agreements with third States or international organisations on security procedures for exchanging and protecting classified information ("security of information agreements");
l'Unione conclude accordi con Stati terzi o organizzazioni internazionali sulle procedure di sicurezza per scambiare e proteggere informazioni classificate ("accordi sulla sicurezza delle informazioni");

the Secretary-General may enter into administrative arrangements on behalf of the GSC in accordance with paragraph 17 of Annex VI where the classification level of EUCI to be released is as a general rule no higher than RESTREINT UE/EU RESTRICTED.
il segretario generale può pattuire intese amministrative a nome dell'SGC

```
conformemente all’allegato VI, punto17, laddove la classifica delle ICUE
da comunicare non supera di norma il livello RESTREINT UE/EU RESTRICTED.
```