LetsMT!

Platform for Online Sharing of Training Data and Building User Tailored MT

www.letsmt.eu/

Project no. 250456

Deliverable D6.4
Evaluation of integration in CAT tools

Version No. 1.0

29/02/2012

Document Information

Deliverable number:	D6.4
Deliverable title:	Evaluation of integration in CAT tools
Due date of deliverable according to DoW:	29/02/2012
Actual submission date of deliverable:	29/02/2012
Main Author(s):	Nansija Lībiete (Tilde), Raivis Skadiņš (Tilde), Josef Šarman (MOR), Tomáš Hudik (MOR)
Participants:	Andrejs Vasiļjevs, Māris Puriņš, Inguna Skadiņa, Indra Sāmīte
Reviewer	Tilde
Workpackage:	WP6
Workpackage title:	MT usage in localisation: facilities and evaluation
Workpackage leader:	MOR
Dissemination Level:	PU
Version:	V1.0
Keywords:	Evaluation, localization, CAT tools

History of Versions

Version	Date	Status	Name of the Author (Partner)	Contributions	Description/Approval Level
0.1	24.01.2012.	draft	Tilde	-	Document structure and description of evaluation methodology prepared.
0.2	22.02.2012.		Tilde	-	Evaluation results from Tilde. Conclusions from Tilde evaluation.
0.3	28.02.2012.		MOR		Evaluation results from Tilde. Conclusions from Tilde evaluation.
1.0	29.02.2012.		Tilde		Summary of Evaluation results.

EXECUTIVE SUMMARY

LetsMT! system integration with CAT tools implemented in Task 6.1 has been evaluated. Industry partners Moravia and Tilde evaluated this application scenario and demonstrated its impact on the software localisation process and professional translators' daily work. Quality, usability, increase of productivity of translation process has been evaluated. Building of domain and project tailored SMT systems for localisation purposes also has been evaluated.

Above mentioned evaluations has been performed one and will be performed at least one more time during the remaining time of the project. Results of the first evaluation will be used to improve translation quality in general as well as MT usability in CAT tools.

Table of Contents

1 Introduction

Growing pressure to reduce translation costs and to increase translation volumes motivates the localization industry to embrace machine translation in addition to other widely used computer assisted translation tools (CAT).

For several decades the most widely used CAT tools in the localization industry have been Translation Memory systems (TM). Since Translation Memories contain fragments of previously translated texts, they can significantly improve the efficiency of localization in cases when the new text is similar to the previously translated material. However, if the text is in a different domain than the TM or in the same domain from a different customer using different terminology, support from the TM is minimal.

The localization industry has experienced increased pressure to provide more efficient and better performing products, particularly due to the fact that volumes of texts that need to be translated are growing at a greater rate than the availability of human translation, and translation results are expected in real-time. For this reason the localization industry is increasingly interested in combining translation memories with machine translation solutions adapted for the particular domain or customer requirements.

Benefits of the application of machine translation in the localization industry are recognized by developers of TM systems. Some developers have already integrated MT in their products or provide such solutions for MT developers. For instance, SDL Trados Studio 2009 supports 3 machine translation engines: SDL Enterprise Translation Server, Language Weaver, and Google Translate. ESTeam TRANSLATOR and Kilgrey's memoQ are other systems providing integration of MT.

For the development of MT in the localization and translation industry, huge pools of parallel texts in a variety of industry formats have been accumulated. The most successful data collection effort is the online repository of TM data by the TAUS Data Association. However, the use of this data alone does not fully utilize the benefits of modern MT technology.

Although the idea to use MT in the localization process is not new, it has not been explored widely in the research community. Different aspects of post-editing and machine translatability have been researched since the nineties (e.g., Berry 1997, Bruckner and Plitt 2001). A comprehensive overview of research on machine translatability and post-editing has been provided by O´Brien (2005). However this work mainly concentrates on the cognitive aspects, not so much on productivity in the localization industry.

Increasing the efficiency of the translation process without a degradation of quality is the most important goal for a localization service provider.

In recent years several productivity tests have been performed in translation and localization industry settings at Microsoft (Schmidtke, 2008), Adobe (Flournoy and Duran, 2009) and Autodesk (Plitt and Masselot, 2010).

The Microsoft Research trained SMT on MS tech domain was used for 3 languages for Office Online 2007 localization: Spanish, French and German. By applying MT to all new words, on average a 5-10% productivity improvement was gained.

In experiments performed by Adobe, about 200,000 words of new text were localized using rule-based MT for translation into Russian (PROMT) and SMT for Spanish and French (Language Weaver). Authors reported an increase of translator's daily output by 22% to 51%.

At Autodesk, a Moses SMT system was evaluated for translation from English to French, Italian, German and Spanish by three translators for each language pair. To measure translation time a special workbench was designed to capture keyboard and pause times for each sentence. Authors reported that although by using MT all translators worked faster, it was in varying proportions: from 20% to 131%. They concluded that MT allowed translators to improve their throughput on average by 74%.

This document describes methodology used for MT evaluation in localization in LetsMT! project and results of an experiment to use for translation SMT integrated into TM in a professional localization company. We present our experiments on the application of an English-Latvian, English-Czech and English-Polish SMT in localization using LetsMT! plug-in into SDL Trados 2009 translation environment. In the localization experiment we measured performance of a translator translating with and without MT. In addition, a quality assessment for texts was performed according to the standard internal quality assessment procedure.

2 Methodology for Evaluation of Machine Translation in Localization

This procedure describes the process and requirements of evaluation of LetsMT! machine translation (MT) in localization scenario.

2.1 Evaluation approach

Evaluation of MT is based on:

1. the measurement of translation performance or productivity,
2. the measurement of translation quality,
3. the time spent for identifying and correcting errors in the translations.

MT systems will be tested against productivity and quality of day-to-day translation using translation memories (TM).

2.2 Scenarios

Translations are performed in SDL Trados Studio 2009 CAT tool environment.

There are 2 scenarios:

1. Translation using TM only (baseline).
2. Translation using TM and MT.

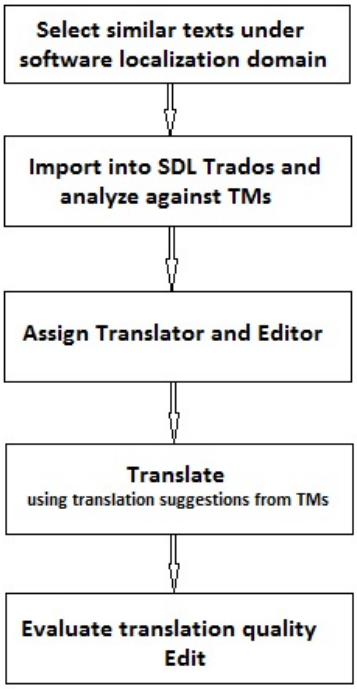

Figure 1. Scenario 1

MT suggestions are provided for every translation unit that does not have a 100% match in TM. Suggestions coming from the MT systems are clearly marked.

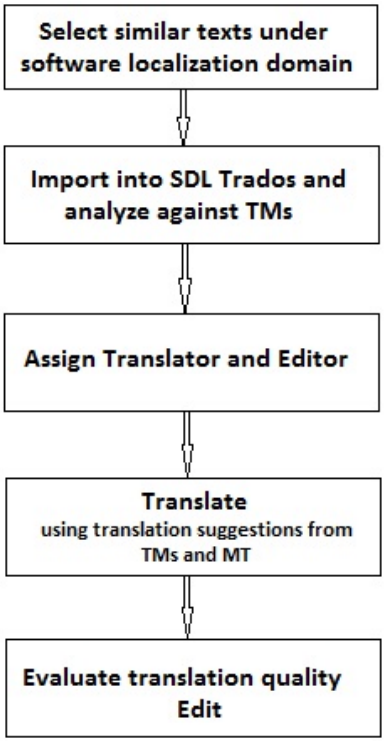

Figure 2. Scenario 2

2.3 Test set selection

Evaluation is made in the software domain for translations from English into target language(s).

Translations for evaluation are selected from texts that have not been translated in the organization before.

The texts (documents) are selected so that ca. 50 % documents contain at least 95% of new words (texts in less used sub-domain, TM does not contain many segments from this sub-domain) and ca. 50% documents contain different fuzzy matches (texts in typical sub-domains, TM contains many segments from this sub-domain).

Texts for tests are used from the following sub-domains of software localization:

- User assistance
- User interface (should not be included in corpora)

There are **2 test sets**: (1) documents in a plain text format without mark-up (formatting or tags), and (2) documents with a mark-up (tags).

Set 1

50 documents without mark-up should be used for Set 1. All documents are split into 2 equal-size parts to perform the two translation scenarios described below. The first part of a document is translated as per scenario 1 and the second part of a document as per scenario 2.

The volume of each part of a document is 500 weighted words on average, resulting in 2 packages of documents with about 25,000 weighted words in each package.

Set 2

10 documents containing text with a mark-up should be used for Set 2. All documents are split into 2 equal-size parts to perform the two translation scenarios described below.

The first part of a document is translated as per scenario 1 and the second part of a document as per scenario 2.

The volume of each part of a document is 500 weighted words on average, resulting in 2 packages of documents with about 5,000 weighted words in each package.

2.4 Test

The evaluation process involves at least 6 translators with different levels of experience and average productivity performance. All translators are well trained to use the MT systems and SDL Trados Studio 2009 in their translation work before measuring their performance in evaluation.

Translators are allowed to use external resources (dictionaries, online reference tools, etc.), just as during regular operations.

Translators perform the test without interruption and switching to other translation tasks on their working day – 8 hours, because splitting the time into short periods would not show trustable performance results. The time spent for translation is reported to the nearest minute.

The first translation made by each translator in scenario 2 should be removed from the result analysis to avoid any “start-up" impact.

Each scenario (scenario 1 and scenario 2) is performed on different working days.

Translators fill in a questionnaire (Appendix 1) when scenario 2 has been completed for each document.

2.4.1 Translation performance and quality assessment

After a document is translated it is evaluated for translation performance and translation quality by editors. The evaluation process involves at least 2 experienced editors. Editors are not aware of the scenario used by the translators. Editors also report their time spent for identifying and correcting errors of the translations and quality assessment to the nearest minute.

There is no inter-editor (inter-annotator) agreement as it is not an everyday practice in localization.

The measurement of translation performance is calculated as a number of weighted words translated per hour. Weighted wordcount is a CAT tool count of words that applies percentage according to various types of matches (new words, fuzzy matches, repetitions, 100% matches). The percentage used in the evaluation is shown in Table 1.

Table 1. The percentage used to calculate weighted wordcount

CAT categories	Count
new words	100%
50% - 74% matches	100%
75% - 84% matches	50%
85% - 94% matches	50%
95% - 99% matches	30%
repetitions	10%
100% matches	10%

Quality of translation is measured by filling in a Quality Assessment (QA) form in accordance with the Tilde QA methodology (*Appendix 2*) based on the industry standard – the Localization Industry Standards Association (LISA) QA model[1]. QA methodology provides a method of measuring the quality of translation. The evaluation process involves inspection of translations and classifying errors according to the following error categories:

- Accuracy
- Language quality
- Style
- Terminology

Preferential changes are not considered as errors.

Performance and quality of work in every of the two translation scenarios is measured and compared for every individual translator. Individual productivity of each translator in the test is measured and compared against his or her own standard productivity. An error score is calculated for every translation task. The error score is a metric calculated by counting errors identified by the editor and applying a weighted multiplier based on the severity of the error type. The error score is calculated per 1,000 weighted words and it is calculated as:

$$ErrorScore = \frac{1000}{n} \sum_{i} w_i e_i$$

where

n is a number of weighted words in a translated text,

e_i is a number of errors of type i,

w_i is a coefficient (weight) indicating severity of type i errors.

There are 15 different error types grouped in 4 error classes: accuracy, language quality, style, and terminology (*Appendix 2*). Different error types influence the error score differently because errors have a different weight depending on the severity of error type.

[1] LISA QA model: http://web.archive.org/web/20080124014404/http://www.lisa.org/products/qamodel/

For example, errors of type comprehensibility (an error that obstructs the user from understanding the information; very clumsy expressions) have weight 3, while errors of type omissions/unnecessary additions have weight 2.

Depending on the error score the translation is assigned a translation quality grade: Superior, Good, Mediocre, Poor, or Very poor (*Table 2*).

Table 2. Quality evaluation based on the score of weighted errors

Error Score	Quality Grade
0…9	Superior
10…29	Good
30…49	Mediocre
50…69	Poor
>70	Very poor

Editors perform quality assessment by marking error categories electronically in the text and filling in a QA form for each translation. Editors inform the project manager when QA is completed.

2.5 Tools

For application in the localization scenario, LetsMT! provides a plug-in for the SDL Trados 2009 CAT environment to use generated MT systems. The MT systems are running on the LetsMT! platform and are accessible using a web service interface based on the SOAP protocol. Connectivity with additional localisation environments will be ensured by providing web services for further integration efforts either by partners or the user community of the LetsMT! service.

The plug-in is has been developed using standard MT integration approach described in SDL Trados SDK. It has been written in .NET (C#), using .NET framework 3.5. The setup is compiled using Nullsoft Install System (NSIS).

To use the plug-in, the user needs to download a setup file from the LetsMT! website (https://www.letsmt.eu/Integration.aspx) and run it. When the user starts SDL TRADOS Studio the plug-in is loaded. Machine translation suggestions from the selected LetsMT! system appears on screen during the translation of the document or can be used to pre-translate documents in the batch process. A SMT system must be specified manually for each language direction.

The baseline scenario establishes the productivity baseline of the current translation process using SDL Trados Studio 2009 when texts are translated unit-by-unit (sentence-by-sentence). The MT scenario measured the impact of using MT in the translation process when translators are provided with not only matches from the translation memory (as in baseline scenario), but also with MT suggestions for every translation unit that does not have a 100% match in translation memory. Suggestions coming from the MT were clearly marked (Figure 3).

We chose to mark MT suggestions clearly because it allows translators to pay more attention to these suggestions. Typically translators trust to suggestions coming from the TM and they make only small changes if it is not a 100% match. Translators are not double-checking terminology, spelling and the grammar of TM suggestions, because the TM contains good quality data. However, translators must pay more attention to suggestions coming from MT, because MT output may be inaccurate, ungrammatical, it may use the wrong terminology, etc.

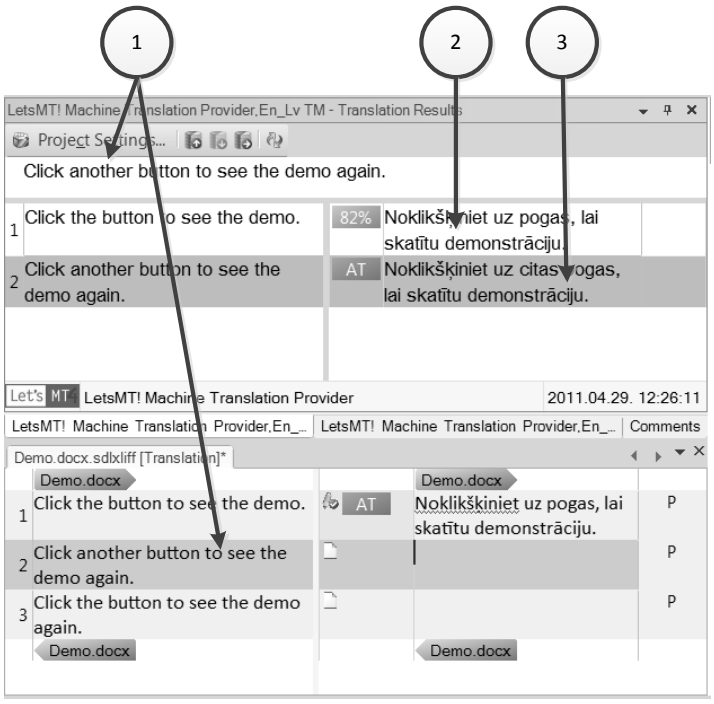

Figure 3. Translation suggestions in SDL Trados Studio 2009; 1 – source text, 2 – a suggestion from the TM, 3 – a suggestion from the MT.

In both scenarios translators were allowed to use whatever external resources needed (dictionaries, online reference tools, etc.), just as during regular operations.

3 Evaluation results

3.1 English-Latvian

3.1.1 SMT system

The total size of the English-Latvian parallel data used to train the translation model is 5.37 M sentence pairs (Table 3).The parallel corpus includes publicly available DGT-TM2 (1.06 M sentences) and OPUS EMEA (0.97 M sentences) corpora (Tiedemann, 2009), as well as a proprietary localization corpus (1.29 M sentences) obtained from translation memories that were created during the localization of interface and user assistance materials for software and user manuals for IT&T appliances. To increase word coverage, word and phrase translations were included from bilingual dictionaries (0.51 M units) from reliable sources with high quality. A larger selection of parallel data was used which was automatically extracted from comparable web corpus (0.9 M sentences) and from 104 works of fiction (0.66 M sentences).

Table 3. Bilingual corpora for English-Latvian system

Bilingual corpus	Parallel units
Localization TM	~1.29 M
DGT-TM	~1.06 M
OPUS EMEA	~0.97 M
Fiction	~0.66 M
Dictionary data	~0.51 M
Web corpus	~0.9 M
Total	**5.37 M**

The monolingual corpus was prepared from news articles from the Web and the monolingual part of the parallel corpora. The total size of the Latvian monolingual corpus was 391 M words (Table 4).

Table 4. Latvian monolingual corpora

Monolingual corpus	Words
Latvian side of parallel corpus	60 M
News (web)	250 M
Fiction	9 M
Total, Latvian	**319 M**

[2] http://langtech.jrc.it/DGT-TM.html

Since Latvian belongs to the class of highly inflected languages with a complex morphology, the SMT system was extended within the Moses (Koehn et al., 2007) framework by integrating morphologic knowledge (Skadiņš et al., 2010). The high inflectional variation of target language increases data sparseness at the boundaries of translated phrases, where a language model over surface forms might be inadequate to estimate the probability of target sentence reliably.

We used the BLEU (Papineni et al., 2002) metric for automatic evaluation. The BLEU score of the SMT system is 35.0 evaluating on a general domain balanced evaluation set and 70.37 evaluating on an IT domain evaluation set. The detailed description of test and development sets and system comparison to other English-Latvian systems are given by Skadiņš et al. (2010).

3.1.2 Test set

The test set for the evaluation was created by selecting documents in the IT domain from the tasks that have not been translated by the translators in the organization before the SMT engine was built. This ensures that translation memories do not contain all the segments of texts used for testing

Documents for translation were selected from the incoming work pipeline if they contained 950-1,050 adjusted words each. Each document was split in half and the first part of it was translated as described in the baseline scenario and the second half of the document – using the MT scenario. The project manager ensured that each part of a single document was translated by different translators so the results are not affected due to translating a familiar document.

Altogether 54 documents were translated. Every document was entered in the translation project tracking system as a separate translation task. An adjusted word is a metric used for quantifying work to be done by translators. Larger documents were split into several fragments.

Although a general purpose SMT system was used, it was trained using specific vendor translation memories as a significant source of parallel corpora. Therefore, the SMT system may be considered slightly biased to a specific IT vendor, or a vendor specific narrow IT domain. The test set contained texts from this vendor and another vendor whose translation memories were not included in the training of the SMT system. We will call these texts as *in narrow IT domain* and *in broad IT domain* for easier reference in the following sections. Approximately 33% of texts translated in each scenario were *in broad IT domain*.

3.1.3 Results

The results were analyzed for 46 translation tasks (23 tasks in each scenario) by analyzing average values for translation performance (translated words per hour) and an error score for translated texts.

Usage of MT suggestions in addition to the use of the translation memories increased productivity of the translators in average from 550 to 731 words per hour (32.9% improvement).

There were significant performance differences in the various translation tasks; the standard deviation of productivity in the baseline and MT scenarios were 213.8 and 315.5 respectively.

At the same time the error score increased for all translators. Although the total increase in the error score was from 20.2 to 28.6 points, it still remained at the quality evaluation grade "Good". We have not performed detailed analysis of reasons causing error score increase yet, but it can be explained by the fact, that translators are tended to trust suggestions coming from the CAT tool and they are not double checking them even if they are marked as MT suggestion.

Grouping of the translation results by narrow/broad domain attribute reveals that MT-assisted translation provides better increase in translation performance for narrow domain (37%) than for broad domain texts (24%). Error scores for both text types are very similar 29.1 and 27.6, respectively.

Grouping of errors identified by error classes reveal the increase of number of errors shown in Table 5.

Table 5. Comparison by error classes, English-Latvian

Error Class	Baseline scenario	MT scenario
Accuracy	6	9
Language quality	6	10
Style	3	4
Terminology	5	7

There were significant differences in the results of different translators from performance increase by 64% to decreased performance by 5% for one of the translators.

Analysis of these differences requires further studies but most likely they are caused by working patterns and the skills of individual translators.

Detailed results of evaluation for English-Latvian are given in Appendix 4.

3.2 English-Polish

3.2.1 SMT System

English-Polish translation engine was trained on 1.5M parallel sentences from Moravia's production data (data of various clients). All the clients were IT companies. The same data was used as a source for monolingual corpus.

The engine was trained without any additional adjustments and parameters, it is a baseline. This means tuning, as well as testing set were filtered out before the training started. Tuning set contained 2000 sentences, while testing set contained 1000 randomly selected sentences (segments). The trained engine achieved: 70.47 BLEU and 0.4812 METEOR score.

3.2.2 Test set

The test set for evaluation of the English – Polish engine was created from Moravia's production data. All the documents belong to IT domain and have not been translated in the organization before.

Segments for translation were taken from real-project data. All documents were divided into fragments with similar size of weighted word count - around 500 words. For every single document half of its fragments were translated as described in the baseline scenario. The remaining fragments were translated using the MT engine. In total 46 fragments were translated.

Even though the MT engine was trained on Moravia production data, most of the testing documents come from broad IT domain (approx. 60%). Client specific translation memories were incorporated in the translation package. So the translators could use inputs from TMs together with MT suggestions.

3.2.3 Results

The results were analyzed for 42 translation tasks (21 tasks in each scenario) by analyzing average values for translation performance (translated words per hour) and an error score for translated texts.

Even though most of the translators reported "poor" quality of MT suggestion the results shows an increased productivity across all documents. The average performance rose from 305 to 392 adjusted words per hour (28.5% improvement).

A significant performance variety has been observed while using MT scenario with 181 words difference compare to 86 under baseline scenario.

Slight decrease of translation quality was recorded. The overall error score increased from 16.8 to 23.6 points. Nevertheless the quality evaluation grade remains "Good". Grouping of errors identified by error classes reveal the increase of number of errors shown in Table 6. Comparison by error classesTable 5. Comparison by error classes, English-Latvian.

Table 6. Comparison by error classes, English-Polish

Error Class	Baseline scenario	MT scenario
Accuracy	2	4
Language quality	1	2
Style	3	4
Terminology	2	3

Results of MT are very sensitive to the training set. The accuracy can be improved by training the engine with more specific data or to have a client dedicated engine.

Language style is the major weakness of automated translations. Even though human translators were supposed to edit the target strings to ensure an appropriate language style is used, Table 6 shows that MT suggestion affected the style in general. Study of this phenomenon and improvement of MT in this area would help to use MT in commercial translations more often.

Detailed results for English-Polish evaluation are attached as Appendix 5.

3.3 English-Czech

3.3.1 SMT System

English-Czech engine was trained on 0.9M sentences. A larger part (1.6M sentences) was taken from Czech National Corpus (topic: **tech domain**) – Institute of Formal and Applied Linguistics (ÚFAL) - http://ufal.mff.cuni.cz/. And the rest (0.5M sentences) were Moravia's production data – different users, all of them were IT companies. LetsMT! filters out duplicate or somehow damaged segments, therefore engine's size (0.9M sentences) is lower than the sum of its constituents.
This is also base-line system, which means that no additional parameters were used. Tuning (2000 sentences) and testing (1000 sentences) were filtered out before the training process has started. After training, tuning and testing took a place. The trained engine achieved: 67.97 BLEU and 0.4668 METEOR score.

3.3.2 Test set

The test set for the evaluation was extracted from Moravia's production data. All the source documents belong to the IT domain and have not been translated in the organization prior the SMT system was trained.

Segments for translation were taken from real-project data. All documents were split into fragments with similar size of weighted word count - around 500 words. For every single document half of its fragments were translated according to the baseline scenario. The remaining part was translated using the MT engine. In total 39 files were translated.

Approximately 70% of the testing content comes from broad IT domain not directly linked with the training data. Therefore client specific translation memories were incorporated in the translation package. Hence the vendors can work with both MT and TM inputs.

3.3.3 Results

The results are based on 34 translation tasks (17 tasks in each scenario) by analyzing average values for translation performance (translated words per hour) and an error score for translated texts.

An increase of productivity by 25.1% was captured while using MT scenario. The average volume of adjusted words per hour rose from 315 to 394.

A quality review discovered minor decrease of translation quality from 19 to 27 error points per 1000 words. Nevertheless the quality evaluation grade is still "Good". Grouping of errors identified by error classes reveal the increase of number of errors shown in Table 7.

Table 7. Comparison by error classes, English-Czech

Error Class	Baseline scenario	MT scenario
Accuracy	4	6
Language quality	1	3
Style	3	3
Terminology	1	2

Detailed analysis of quality degradation might be subject of another study.

Despite the fact that 59% of Czech translators evaluated generally the MT input as inefficient, the evaluation results shows that use of MT system significantly contributes to increase the translation productivity.

Results of MT are very sensitive to the training set. The accuracy can be improved by training the engine with more specific data or to have a client dedicated engine.

Language style is the major weakness of automated translations. Even though human translators were supposed to edit the target strings to ensure an appropriate language style is used, Table 7 shows that MT suggestion affected the style in general. Study of this phenomenon and improvement of MT in this area would help to use MT in commercial translations more often.

Detailed results of English-Czech evaluation are provided below as Appendix 6.

4 Conclusions and Future work

Current development of SMT tools and techniques in LetsMT! project has reached the level where they can be implemented in practical applications addressing the needs of large user groups in a variety of application scenarios.

Results promise important advances in the application of SMT in localization by integrating available tools and technologies into an easy-to-use cloud-based platform for data sharing and generation of customized MT. Building of domain and project tailored SMT systems for localisation purposes has been evaluated and results show that the current LetsMT! platform allows to train SMT systems which are practically usable in localization and help to increase translator productivity.

The results of our experiment clearly demonstrate that it is feasible to integrate the current state of the art SMT systems for highly inflected languages into the localization process.

The use of the English->Latvian SMT suggestions in addition to the translation memories in the SDL Trados CAT tool lead to the increase of translation performance by 32.9% while maintaining an acceptable quality of translation. Even better performance results are achieved when using a customized SMT system that is trained on a specific domain and/or same customer parallel data.

Error rate analysis shows that overall usage of MT suggestions decrease the quality of the translation in all error categories, particularly in language quality. At the same time this degradation is not critical and the result is acceptable for production purposes.

The evaluation of English->Polish and English->Czech MT engine confirmed positive results reached during English->Latvian tests. Despite minor quality issues the increase of translation performance by 28.5% and 25.1%, respectively proves SMT system to have considerable impact to the translation productivity and related localization costs.

Results of the evaluation are summarized in Table 8.

In the future we are going to perform this experiment on larger scale. We will repeat similar experiments (i) involving more translators, (ii) translating texts in different domains, (iii) in other language pairs, and (iv) evaluating translation of texts with a mark-up (formatting, hyperlinks, references and other tags). More detailed analysis of reasons causing error score increase in MT scenario also will be performed.

Table 8. Summary of evaluation results

	English-Latvian	English-Polish	English-Czech
MT training data			
parallel sentences	5.37 M	1.5M	0.9M
monolingual sentences	319 M/words	1.5M	0.9M
BLEU score	70.37	70.47	67.97
METEOR score	0.4807	0.4812	0.4668
Baseline scenario			
words per hour	550	305	315
error score	20.2	16.8	19.0
MT scenario			
words per hour	731	392	394
error score	28.6	23.6	27.0
Productivity increase	32.9 %	28.5 %	25.1 %

References

Berry Mark. 1997. Integrating Trados translator's workbench with Machine Translation. *Proceedings of Machine Translation Summit VI.*

Bruckner, Christine and Mirko Plitt. 2001. Evaluating the operational benefit of using machine translation output as translation memory input. *MT Summit VIII, MT evaluation: who did what to whom (Fourth ISLE workshop)*, 61–65.

Flournoy, Raymond and Christine Duran. 2009. Machine translation and document localization at Adobe: From pilot to production. *MT Summit XII: Proceedings of the Twelfth Machine Translation Summit*, Ottawa, Canada.

Koehn, Philipp, Federico M., Cowan B., Zens R., Duer C., Bojar O., Constantin A., Herbst E. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. *Proceedings of the ACL 2007 Demo and Poster Sessions*, Prague, 177-180.

O´Brien, Sharon. 2005. Methodologies for measuring the correlations between post-editing effort and machine translatability. *Machine Translation*, 19(1):37–58, March 2005.

Papineni K., Roukos S., Ward T., Zhu W. 2002. BLEU: a method for automatic evaluation of machine translation, *Proceedings of the 40th Annual Meeting of the Association of Computational Linguistics (ACL).*

Plitt Mirko, François Masselot. 2010. A Productivity Test of Statistical Machine Translation Post-Editing in a Typical Localisation Context. *The Prague Bulletin of Mathematical Linguistics*, 93(January 2010): 7–16

Schmidtke, Dag. 2008. Microsoft office localization: use of language and translation technology.

Skadiņš, Raivis, Kārlis Goba and Valters Šics. 2010. Improving SMT for Baltic Languages with Factored Models. *Proceedings of the Fourth International Conference Baltic HLT 2010*, Riga.

Tiedemann, Jörg. 2009. News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces. *N. Nicolov, K. Bontcheva, G. Angelova, R. Mitkov (eds.) Recent Advances in Natural Language Processing* (vol V), John Benjamins, Amsterdam/Philadelphia, 237-248.

Appendix 1

Questionnaire_Sc2_[filename]

DOCUMENT TITLE: [enter the file name here]

QUESTION: What was the average quality of MT suggestions in cases where text was not found in TM (100% match)?

3. Good (minimal changes needed)

2. Average (MT suggested translation was useful but had to be edited)

1. Poor (MT suggestion was misleading or translation from TM (non-100% match) was used, or translation "from scratch" was more efficient)

ANSWER: [enter the number here]

Appendix 2

Tilde Localization QA form - Translation Quality Assessment

This form is filled out by an Editor or a Language Specialist.

Please see procedural notes and description of error categories in **Error categories** sheet.

Fill in the **Basic information** section, **Amount of errors** column and **General comment** field.

Basic information	
Project name:	
File name:	
Source language:	
Target language:	
Translator:	
Validated by:	
Validation date:	
Stylistic type (please, select):	
Number of words checked:	1000

Error Category	Weight	Amount of errors	Negative points
1. Accuracy			
1.1. Understanding of the source text	3		0
1.2. Understanding the functionality of the product	3		0
1.3. Comprehensibility	3		0
1.4. Omissions/Unnecessary additions	2		0
1.5. Translated/Untranslated	1		0
1.6. Left-overs	1		0
Total			**0**
2. Language quality			
2.1. Grammar	2		0
2.2. Punctuation	1		0
2.3. Spelling	1		0
Total			**0**
3. Style			
3.1. Word order, word-for-word translation	1		0
3.2. Vocabulary and style choice	1		0
3.3. Style Guide adherence	2		0
3.4. Country standards	1		0
Total			**0**
4. Terminology			
4.1. Glossary adherence	2		0

4.2. Consistency	2		0
Total			**0**
Grand Total			**0**
Error Score (negative points) per 1000 words			**0**
Quality:			**Superior**

General comment:

Final assessment is done as follows:	**Score scale**	
Negative points for errors of each category are calculated according to the formula:	**Error score**	**Quality grade**
"Number of errors of given type" x "Error weight"	0…9	Superior
Weighted score is calculated according to the following formula:	10…29	Good
(Total negative points / Wordcount) x 1000	30…49	Mediocre
Final quality assessment is done according to the **Score Scale.**	50…69	Poor
	70…	Very poor

Notes:

In case of recurring errors (double space, the same spelling or terminology error) they should only be counted once.

Each error is counted once, by the most appropriate category. If in doubt, use the first appropriate category (top-down).

Preferential changes should not be counted as negative points, but they may be listed in a separate Comments spreadsheet.

Category	Description
Accuracy	
Understanding of the source text	A lack of comprehension of the source text resulting in incorrect meaning of the translation.
Understanding the functionality of the product	Translation does not comply with the actual function of the product. The translation of the word is OK as such but incorrect in the context.
Comprehensibility	Any error that obstructs the user from understanding the information. Very clumsy expressions.
Omissions/unnecessary additions	Words, part of sentences, sentences, paragraphs are missing. No relevant information in the source language should be omitted in the translation, unless specifically requested. The translation should not contain any unnecessary text.
Translated/Untranslated	Parts that were supposed to be translated were not translated or parts that should not be translated were translated.
Left-overs	Redundant words resulting from sentence change, wrong declinations resulting from correcting one word only but not the rest. Unnecessary question marks or asterisks left in translated text.
Language quality	
Grammar	Grammar, syntax or morphology rules are broken.
Punctuation	Incorrect usage of punctuation marks - full stops missing, opening or closing punctuation marks (quote, parenthesis), double spaces, etc.
Spelling	The translation should contain no spelling errors.
Style	
Word order, word-for-word translation	Functional sentence perspective (theme, rheme), word order. Word for word translation, resulting in stylistically inappropriate expression.
Vocabulary and style choice	Archaisms, jargon, colloquial words, verbosity, inappropriate style.
Style Guide adherence	Product Style Guide rules are ignored. In case of absence of Product Style Guide definite company style rules must be observed. Standard phrases must be used - in case of technical documentation.
Country standards	Adaptation of country standards (date and time formats, units of measurement, currency, number formats, sorting order, capitalization etc.). Examples (of names, streets, etc.) are not localized.
Terminology	
Glossary adherence	Translation does not adhere to the terms in the glossary of project/product, or does not use generally available industry terminology. Technical documentation does not use the correct translation of interface elements.
Consistency	Inconsistent usage of translation for one term or title (for cross-references).

Quality Assessment form, Values for form fields

Yes/No	Yes No
Languages	English Estonian Latvian Lithuanian
Text Type	User interface User assistance, tech. documentation Medicine Legal Marketing or Web material
Quality	Superior Good Mediocre Poor Very poor
Error category	Accuracy Language quality Style Terminology Preferential

Appendix 3

Form: Summary of the evaluation results

Task ID	Translator name	Translator qualification	Estimated time (h)	Planned performance	Actual time	Actual performance	Quality assessment (Appendix 2)						Quality grade	MT quality (References)	Editor name	Editing estimated time	Editing planned performance	Editing actual time	Editing actual performance
		(translator, senior translator)	h	(weighted words/h)	h	(weighted words/h)	Accuracy	Language quality	Style	Terminology	Count of weighted errors	Error score (total (per 1000 weighted words)	(Superior, Good, Mediocre, Poor, Very Poor)	Score 1-3 (where 3 – the best)		h	(total words/h)	h	(total words/h)

Appendix 4

Detailed results of evaluation for English-Latvian

Task ID	Scenario,	Text size,	Text origin,	Translator name	Translator qualification	Estimated time	Planned performance,	Actual time	Actual performance,	Quality assesment, negative points						Quality total valuation	MT quality feedback
	(S1, S2)	(adjusted words)				h	(adjusted words/h)	h	(adjusted words/h)	Accuracy	Language quality	Style	Terminology	Total	ErrorScore	(Superior, Good, Mediocre, Poor, Very Poor)	Score 1-3(best)
Sc1_Tr4_D13-1 (	S1	486,6	Client1	Dana Grīnbauma	Translator	1,39	350	1,00	487	3	4	0	2	9	17	Good	n/a
Sc1_Tr4_D14-1 (	S1	484	Client1	Dana Grīnbauma	Translator	1,38	350	0,80	605	5	2	0	2	9	19	Good	n/a
Sc1_Tr3_D6-1 (A	S1	512	Client1	Artūrs Pudulis	Translator	1,46	350	1,50	341	0	5	0	2	7	14	Good	n/a
Sc1_Tr3_D7-1 (A	S1	507	Client1	Artūrs Pudulis	Translator	1,45	350	1,30	390	0	3	2	2	7	14	Good	n/a
Sc1_Tr5_D16-1 (	S1	466,5	Client1	Mārtiņš Kore	Translator	1,33	350	1,25	373	8	4	2	2	16	33	Mediocre	n/a
Sc1_Tr5_D17-1 (	S1	497,1	Client1	Mārtiņš Kore	Translator	1,42	350	1,40	355	7	2	2	4	15	28	Good	n/a
Sc1_Tr2_D3-1 (J	S1	482,1	Client1	Jānis Šlapiņš	Senior Transla	1,38	350	1,00	482	5	1	0	8	14	28	Good	n/a
Sc1_Tr2_D4-1 (J	S1	490,4	Client1	Jānis Šlapiņš	Senior Transla	1,40	350	1,10	446	9	6	3	4	22	42	Mediocre	n/a
Sc1_Tr1_D1-1 (J	S1	522	Client1	Juris Celmiņš	Senior Transla	1,49	350	0,57	916	1	2	1	4	8	15	Good	n/a
Sc1_Tr1_D2-1 (J	S1	496	Client1	Juris Celmiņš	Senior Transla	1,42	350	0,62	800	4	2	0	0	6	12	Good	n/a
Sc1_Tr1_D10-1 (	S1	511,4	Client2	Juris Celmiņš	Senior Transla	1,46	350	1,05	487	3	2	0	0	5	10	Superior	n/a
Sc1_Tr1_D10-2 (	S1	490,8	Client2	Juris Celmiņš	Senior Transla	1,40	350	0,80	614	0	3	2	2	7	14	Good	n/a
Sc1_Tr2_D8-1 (J	S1	496,3	Client2	Jānis Šlapiņš	Senior Transla	1,42	350	1,40	355	2	2	0	6	10	19	Good	n/a
Sc1_Tr2_D9-1 (J	S1	496,3	Client2	Jānis Šlapiņš	Senior Transla	1,42	350	1,00	496	0	4	0	0	4	8	Superior	n/a
Sc1_Tr3_D18-1 (	S1	464,4	Client2	Artūrs Pudulis	Translator	1,33	350	0,90	516	0	4	2	2	8	16	Good	n/a
Sc1_Tr3_D19-1 (	S1	454,5	Client2	Artūrs Pudulis	Translator	1,30	350	1,00	455	2	8	3	4	17	36	Mediocre	n/a
Sc1_Tr1_D15-1 (	S1	509	Client3	Juris Celmiņš	Senior Transla	1,45	350	0,58	878	0	3	1	2	6	12	Good	n/a
Sc1_Tr1_D15-3 (	S1	492	Client3	Juris Celmiņš	Senior Transla	1,41	350	0,52	946	0	3	2	2	7	14	Good	n/a
Sc1_Tr1_D15-5 (	S1	489	Client3	Juris Celmiņš	Senior Transla	1,40	350	0,47	1040	4	2	4	0	10	20	Good	n/a
Sc1_Tr2_D12-5 (	S1	518,3	Client3	Jānis Šlapiņš	Senior Transla	1,48	350	1,25	415	3	0	1	0	4	8	Superior	n/a
Sc1_Tr2_D12-7 (	S1	505	Client3	Jānis Šlapiņš	Senior Transla	1,44	350	1,40	361	4	0	5	4	13	26	Good	n/a
Sc1_Tr2_D12-9 (	S1	479	Client3	Jānis Šlapiņš	Senior Transla	1,37	350	1,25	383	0	4	2	6	12	25	Good	n/a
Sc1_Tr5_D12-3 (	S1	501	Client3	Mārtiņš Kore	Translator	1,43	350	1,00	501	11	2	2	2	17	34	Mediocre	n/a
		11351				**1,41**	**350,0**	**1,01**	**549,6**	**3,1**	**3,0**	**1,5**	**2,6**	**10,1**	**20,2**		
		total				avg	avg	avg	avg	avg	avg	avg	avg	avg	avg		
								23,16							***Good***		
								total									

Task ID	Scenario,	Text size,	Text origin,	Translator name	Translator qualification	Estimated time	Planned performance,	Actual time	Actual performance,	Quality assesment, negative points						Quality total valuation	MT quality feedback
	(S1, S2)	(adjusted words)				h	(adjusted words/h)	h	(adjusted words/h)	Accuracy	Language quality	Style	Terminology	Total	ErrorScore	(Superior, Good, Mediocre, Poor, Very Poor)	Score 1-3(best)
Sc2_Tr4_D13-2 (	S2	487	Client1	Dana Grīnbauma	Translator	1,39	350	1,10	443	2	8	3	6	19	37	Mediocre	2
Sc2_Tr4_D14-2 (	S2	474,4	Client1	Dana Grīnbauma	Translator	1,36	350	1,10	431	0	4	3	4	11	20	Good	2
Sc2_Tr4_D16-1 (	S2	466,5	Client1	Dana Grīnbauma	Translator	1,33	350	0,60	778	6	7	1	0	14	29	Good	3
Sc2_Tr3_D20-1 (	S2	484	Client1	Artūrs Pudulis	Translator	1,38	350	0,80	605	8	3	1	4	16	31	Mediocre	2
Sc2_Tr3_D20-2 (	S2	474,4	Client1	Artūrs Pudulis	Translator	1,36	350	0,55	863	6	12	0	6	24	44	Mediocre	2
Sc2_Tr1_D6-2 (J	S2	505	Client1	Juris Celmiņš	Senior Transla	1,44	350	0,55	918	6	5	6	0	17	34	Mediocre	2
Sc2_Tr1_D7-2 (J	S2	513,2	Client1	Juris Celmiņš	Senior Transla	1,47	350	0,43	1193	1	4	0	2	7	14	Good	2
Sc2_Tr2_D2-2 (J	S2	529,5	Client1	Jānis Šlapiņš	Senior Transla	1,51	350	1,20	441	3	0	3	2	8	14	Good	2
Sc2_Tr5_D17-2 (	S2	543,5	Client1	Mārtiņš Kore	Translator	1,55	350	1,15	473	7	10	1	0	18	32	Mediocre	2
Sc2_Tr1_D8-2 (J	S2	461	Client2	Juris Celmiņš	Senior Transla	1,32	350	0,52	887	0	4	0	10	14	28	Good	2
Sc2_Tr1_D9-2 (J	S2	496,3	Client2	Juris Celmiņš	Senior Transla	1,42	350	0,43	1154	2	6	3	2	13	26	Good	2
Sc2_Tr2_D8-2 (J	S2	461	Client2	Jānis Šlapiņš	Senior Transla	1,32	350	1,00	461	4	2	0	0	6	12	Good	2
Sc2_Tr2_D9-2 (J	S2	513	Client2	Jānis Šlapiņš	Senior Transla	1,47	350	1,15	446	0	0	1	6	7	14	Good	2
Sc2_Tr3_D18-2 (	S2	455,5	Client2	Artūrs Pudulis	Translator	1,30	350	0,65	701	12	5	6	10	33	65	Poor	2
Sc2_Tr3_D19-2 (	S2	476,1	Client2	Artūrs Pudulis	Translator	1,36	350	0,80	595	3	6	4	6	19	37	Mediocre	1
Sc2_Tr1_D15-2 (	S2	491	Client3	Juris Celmiņš	Senior Transla	1,40	350	0,60	818	2	2	2	4	10	20	Good	1
Sc2_Tr1_D15-4 (	S2	486	Client3	Juris Celmiņš	Senior Transla	1,39	350	0,37	1314	0	7	2	2	11	23	Good	2
Sc2_Tr1_D15-6 (	S2	486	Client3	Juris Celmiņš	Senior Transla	1,39	350	0,33	1473	1	4	1	2	8	16	Good	2
Sc2_Tr2_D12-10	S2	501,1	Client3	Jānis Šlapiņš	Senior Transla	1,43	350	1,15	436	6	11	1	0	18	35	Mediocre	1
Sc2_Tr2_D12-6 (	S2	514	Client3	Jānis Šlapiņš	Senior Transla	1,47	350	1,45	354	0	3	4	2	9	18	Good	1
Sc2_Tr2_D12-8 (	S2	520	Client3	Jānis Šlapiņš	Senior Transla	1,49	350	1,10	473	0	2	2	2	6	12	Good	1
Sc2_Tr5_D12-2 (	S2	505	Client3	Mārtiņš Kore	Translator	1,44	350	0,60	842	16	3	0	4	23	46	Mediocre	2
Sc2_Tr5_D12-4 (	S2	495	Client3	Mārtiņš Kore	Translator	1,41	350	0,70	707	13	6	2	4	25	51	Poor	2
		11339				**1,41**	**350,00**	**0,80**	**730,6**	**4,3**	**5,0**	**2,0**	**3,4**	**14,6**	**28,6**		**1,8**
		total				avg	avg	avg	avg	avg	avg	avg	avg	avg	avg		
															Good		
								18,33									
								total									

Appendix 5

Detailed results of evaluation for English-Polish

Task ID	Scenario	Text size	Text origin	Translator name	Translator qualification	Estimated time (h)	Planned performance	Actual time	Actual performance	Quality assessment (Appendix 2)						Quality grade	MT quality (References)
	S1,S2	(adjusted words)			(translator, senior translator)	h	(weighted words/h)	h	(weighted words/h)	Accuracy	Language quality	Style	Terminology	Count of weighted errors	Error score (total (per 1000 weighted words)	(Superior, Good, Mediocre, Poor, Very Poor)	Score 1-3 (where 3 – the best)
A_PL_01-1	S1	456,2	cust-A	Andrzej Sawicki	Translator	1,63	280	1,45	315	1	0	7	2	10	22	Good	-
A_PL_01-2	S1	491,4	cust-A	Andrzej Sawicki	Translator	1,76	280	1,57	313	1	1	1	2	5	10	Good	-
A_PL_01-5	S1	472,1	cust-A	Andrzej Sawicki	Translator	1,69	280	1,50	315	5	0	3	4	12	25	Good	-
A_PL_01-7	S1	542,5	cust-A	Maksymilian Nawrocki	Senior Translator	1,94	280	1,75	310	2	2	4	0	8	15	Good	-
A_PL_02-1	S1	481,1	cust-A	Maksymilian Nawrocki	Senior Translator	1,72	280	1,40	344	0	1	2	0	3	6	Superior	-
A_PL_03-1	S1	505	cust-A	Maksymilian Nawrocki	Senior Translator	1,80	280	1,53	330	0	2	2	2	6	12	Good	-
A_PL_04-2	S1	522,4	cust-B	Agata Reszke	Senior Translator	1,87	280	1,73	302	0	1	3	2	6	11	Good	-
A_PL_05-2	S1	514	cust-A	Maksymilian Nawrocki	Senior Translator	1,84	280	1,72	299	2	1	0	2	5	10	Good	-
A_PL_06-1	S1	466,1	cust-B	Maksymilian Nawrocki	Senior Translator	1,66	280	1,55	301	3	0	2	0	5	11	Good	-
A_PL_06-3	S1	472,5	cust-B	Maksymilian Nawrocki	Senior Translator	1,69	280	1,57	301	1	2	0	0	3	6	Superior	-
A_PL_06-5	S1	562,3	cust-B	Agata Reszke	Senior Translator	2,01	280	1,85	304	4	1	2	4	11	20	Good	-
A_PL_07-2	S1	487	cust-C	Andrzej Sawicki	Translator	1,74	280	1,53	318	3	2	2	2	9	18	Good	-
A_PL_07-3	S1	541,2	cust-C	Andrzej Sawicki	Translator	1,93	280	1,75	309	2	3	3	0	8	15	Good	-
A_PL_10-1	S1	529,1	cust-C	Andrzej Sawicki	Translator	1,89	280	1,68	315	4	2	1	0	7	13	Good	-
A_PL_10-2	S1	534	cust-C	Maksymilian Nawrocki	Senior Translator	1,91	280	1,72	310	3	0	3	2	8	15	Good	-
A_PL_10-4	S1	492,5	cust-C	Maksymilian Nawrocki	Senior Translator	1,76	280	1,60	308	0	0	0	0	0	0	Superior	-
A_PL_11-1	S1	467,2	cust-C	Andrzej Sawicki	Translator	1,67	280	1,48	316	5	1	3	2	11	24	Good	-
A_PL_09-1	S1	503,1	cust-C	Witold Grzebinski	Translator	1,80	280	1,93	261	4	3	6	4	17	34	Mediocre	-
A_PL_09-3	S1	507,5	cust-C	Witold Grzebinski	Translator	1,81	280	1,95	260	6	1	8	6	21	41	Mediocre	-
A_PL_12-1	S1	536	cust-D	Agata Reszke	Senior Translator	1,91	280	1,68	319	2	2	3	2	9	17	Good	-
A_PL_13-1	S1	482,5	cust-D	Witold Grzebinski	Translator	1,72	280	1,87	258	3	2	4	4	13	27	Good	-
Total		**10565,7**				**1,80**	**280,00**	**1,66**	**305,14**	**2,43**	**1,29**	**2,81**	**1,90**	**8,43**	**16,76**	**Good**	

Task ID	Scenario	Text size	Text origin	Translator name	Translator qualification	Estimated time (h)	Planned performance	Actual time	Actual performance	Quality assessment (Appendix 2)						Quality grade	MT quality (References)
	S1,S2	(adjusted words)			(translator, senior translator)	h	(weighted words/h)	h	(weighted words/h)	Accuracy	Language quality	Style	Terminology	Count of weighted errors	Error score (total (per 1000 weighted words)	(Superior, Good, Mediocre, Poor, Very Poor)	Score 1-3 (where 3 – the best)
B_PL_01-3	S2	488,2	cust-A	Agata Reszke	Senior Translator	1,74	280	0,97	503	1	2	5	0	8	16	Good	2
B_PL_01-4	S2	518	cust-A	Andrzej Sawicki	Translator	1,85	280	1,23	421	1	0	5	2	8	15	Good	1
B_PL_01-6	S2	505,2	cust-A	Maksymilian Nawrocki	Senior Translator	1,80	280	1,23	411	0	0	3	2	5	10	Good	1
B_PL_02-2	S2	512,4	cust-A	Maksymilian Nawrocki	Senior Translator	1,83	280	1,25	410	1	0	2	1	4	8	Superior	1
B_PL_04-1	S2	480,5	cust-B	Agata Reszke	Senior Translator	1,72	280	1,18	407	6	0	2	4	12	25	Good	1
B_PL_05-1	S2	489,2	cust-A	Maksymilian Nawrocki	Senior Translator	1,75	280	1,06	462	0	0	2	4	6	12	Good	2
B_PL_05-3	S2	506,2	cust-A	Maksymilian Nawrocki	Senior Translator	1,81	280	1,15	440	2	4	2	4	12	24	Good	2
B_PL_06-2	S2	473,5	cust-B	Maksymilian Nawrocki	Senior Translator	1,69	280	1,12	423	2	3	3	2	10	21	Good	1
B_PL_06-4	S2	502,1	cust-B	Agata Reszke	Senior Translator	1,79	280	1,20	418	1	2	5	2	10	20	Good	1
B_PL_07-1	S2	479	cust-C	Andrzej Sawicki	Translator	1,71	280	1,30	368	2	0	0	4	6	13	Good	1
B_PL_07-4	S2	509,8	cust-C	Andrzej Sawicki	Translator	1,82	280	1,35	378	3	4	3	4	14	27	Good	1
B_PL_08-1	S2	499,4	cust-C	Andrzej Sawicki	Translator	1,78	280	1,40	357	6	3	8	4	21	42	Mediocre	1
B_PL_10-6	S2	533,1	cust-C	Agata Reszke	Senior Translator	1,90	280	1,18	452	6	0	5	4	15	28	Good	2
B_PL_10-3	S2	486	cust-C	Andrzej Sawicki	Translator	1,74	280	1,37	355	4	1	6	2	13	27	Good	1
B_PL_10-5	S2	541,4	cust-C	Andrzej Sawicki	Translator	1,93	280	1,45	373	6	3	2	0	11	20	Good	1
B_PL_08-2	S2	453,4	cust-C	Andrzej Sawicki	Translator	1,62	280	1,30	349	7	4	4	0	15	33	Mediocre	1
B_PL_09-2	S2	487,2	cust-C	Andrzej Sawicki	Translator	1,74	280	1,38	353	8	1	12	4	25	51	Poor	1
B_PL_09-4	S2	496,1	cust-C	Andrzej Sawicki	Translator	1,77	280	1,42	349	5	2	3	2	12	24	Good	1
B_PL_12-2	S2	513	cust-D	Witold Grzebinski	Translator	1,83	280	1,53	335	2	2	2	6	12	23	Good	1
B_PL_13-2	S2	507,6	cust-D	Witold Grzebinski	Translator	1,81	280	1,50	338	7	2	4	4	17	33	Mediocre	1
B_PL_14-1	S2	579	cust-D	Witold Grzebinski	Translator	2,07	280	1,80	322	5	1	4	4	14	24	Good	2
Total		**10560**				**1,80**	**280,00**	**1,30**	**391,62**	**3,57**	**1,62**	**3,90**	**2,81**	**11,90**	**23,62**	**Good**	**1,24**

Appendix 6

Detailed results of evaluation for English-Czech

Task ID	Scenario	Text size	Text origin	Translator name	Translator qualification	Estimated time (h)	Planned performance	Actual time	Actual performance	Quality assessment (Appendix 2)						Quality grade	MT quality (References)
	S1,S2	(adjusted words)			(translator, senior translator)	h	(weighted words/h)	h	(weighted words/h)	Accuracy	Language quality	Style	Terminology	Count of weighted errors	Error score (total (per 1000 weighted words)	(Superior, Good, Mediocre, Poor, Very Poor)	Score 1-3 (where 3 – the best)
A_CZ_01-1	S1	527	cust-A	Barbora Zlama	Translator	1,88	280	1,75	301	4	0	3	2	9	17	Good	-
A_CZ_01-3	S1	511,4	cust-A	Milan Vesely	Translator	1,83	280	1,58	324	2	1	3	2	8	16	Good	-
A_CZ_03-1	S1	569	cust-A	Jan Trhlik	Translator	2,03	280	1,95	292	5	4	6	4	19	33	Mediocre	-
A_CZ_04-2	S1	501,7	cust-A	Milan Vesely	Translator	1,79	280	1,48	339	1	2	3	0	6	12	Good	-
A_CZ_04-4	S1	504,4	cust-A	Barbora Zlama	Translator	1,80	280	1,62	311	5	2	3	0	10	20	Good	-
A_CZ_05-1	S1	547,2	cust-A	Ales Horak	Senior Translat	1,95	280	1,43	383	3	0	3	1	7	13	Good	-
A_CZ_05-3	S1	552	cust-A	Ales Horak	Senior Translat	1,97	280	1,33	415	3	0	1	0	4	7	Superior	-
A_CZ_06-2	S1	574,5	cust-B	Daniela Skotni	Senior Translat	2,05	280	1,72	334	2	0	1	2	5	9	Superior	-
A_CZ_06-4	S1	549,8	cust-B	Daniela Skotni	Senior Translat	1,96	280	1,67	329	3	1	2	0	6	11	Good	-
A_CZ_06-6	S1	477,1	cust-B	Jan Trhlik	Translator	1,70	280	1,93	247	9	6	8	2	25	52	Poor	-
A_CZ_07-2	S1	452	cust-B	Jan Trhlik	Translator	1,61	280	1,78	254	4	4	8	2	18	40	Mediocre	-
A_CZ_09-1	S1	554,6	cust-B	Milan Vesely	Translator	1,98	280	1,80	308	2	3	6	0	11	20	Good	-
A_CZ_09-3	S1	561,5	cust-B	Milan Vesely	Translator	2,01	280	1,73	325	1	1	2	2	6	11	Good	-
A_CZ_09-5	S1	471,3	cust-B	Milan Vesely	Translator	1,68	280	1,48	318	2	1	3	1	7	15	Good	-
A_CZ_10-2	S1	560,2	cust-C	Ales Horak	Senior Translat	2,00	280	1,82	308	4	0	2	0	6	11	Good	-
A_CZ_10-4	S1	509,1	cust-C	Barbora Zlama	Translator	1,82	280	1,98	257	9	0	4	2	15	29	Good	-
A_CZ_11-2	S1	518	cust-C	Daniela Skotni	Senior Translat	1,85	280	1,67	310	2	0	1	0	3	6	Superior	-
Total		**8940,8**				**1,88**	**280,00**	**1,69**	**315,00**	**3,59**	**1,47**	**3,47**	**1,18**	**9,71**	**18,94**	**Good**	

Task ID	Scenario	Text size	Text origin	Translator name	Translator qualification	Estimated time (h)	Planned performance	Actual time	Actual performance	Quality assessment (Appendix 2)						Quality grade	MT quality (References)
	S1,S2	(adjusted words)			(translator, senior translator)	h	(weighted words/h)	h	(weighted words/h)	Accuracy	Language quality	Style	Terminology	Count of weighted errors	Error score (total (per 1000 weighted words)	(Superior, Good, Mediocre, Poor, Very Poor)	Score 1-3 (where 3 – the best)
B_CZ_01-2	S2	505	cust-A	Barbora Zlama	Translator	1,80	280	1,38	366	8	2	2	0	12	24	Good	1
B_CZ_02-1	S2	463,5	cust-A	Daniela Skotni	Senior Translat	1,66	280	0,90	515	2	2	3	4	11	24	Good	2
B_CZ_04-1	S2	495,2	cust-A	Milan Vesely	Translator	1,77	280	1,22	406	10	6	1	2	19	38	Mediocre	1
B_CZ_04-3	S2	517	cust-A	Milan Vesely	Translator	1,85	280	1,13	458	7	0	4	0	11	21	Good	2
B_CZ_04-5	S2	484,1	cust-A	Barbora Zlama	Translator	1,73	280	1,25	387	11	4	7	2	24	50	Poor	2
B_CZ_05-2	S2	493	cust-A	Ales Horak	Senior Translat	1,76	280	0,98	503	3	1	2	2	8	16	Good	2
B_CZ_06-1	S2	481,2	cust-B	Daniela Skotni	Senior Translat	1,72	280	1,12	430	4	0	4	4	12	25	Good	1
B_CZ_06-3	S2	469	cust-B	Daniela Skotni	Senior Translat	1,68	280	1,08	434	3	0	1	2	6	13	Good	1
B_CZ_06-5	S2	480,3	cust-B	Daniela Skotni	Senior Translat	1,72	280	1,17	411	8	0	2	2	12	25	Good	2
B_CZ_07-1	S2	521	cust-B	Jan Trhlik	Translator	1,86	280	1,78	293	5	9	4	6	24	46	Mediocre	2
B_CZ_08-1	S2	495	cust-B	Daniela Skotni	Senior Translat	1,77	280	1,28	387	8	2	1	0	11	22	Good	1
B_CZ_09-2	S2	563,2	cust-B	Milan Vesely	Translator	2,01	280	1,57	359	2	7	0	2	11	20	Good	1
B_CZ_09-4	S2	525,7	cust-B	Milan Vesely	Translator	1,88	280	1,53	344	10	6	4	0	20	38	Mediocre	1
B_CZ_10-1	S2	474	cust-C	Ales Horak	Senior Translat	1,69	280	1,13	419	7	0	2	0	9	19	Good	2
B_CZ_10-3	S2	500,5	cust-C	Barbora Zlama	Translator	1,79	280	1,68	298	14	2	4	4	24	48	Mediocre	1
B_CZ_11-1	S2	572,5	cust-C	Daniela Skotni	Senior Translat	2,04	280	1,62	353	3	1	1	0	5	9	Superior	1
B_CZ_11-3	S2	489	cust-C	Daniela Skotni	Senior Translat	1,75	280	1,43	342	4	2	2	2	10	20	Good	1
Total		**8529,2**				**1,79**	**280,00**	**1,31**	**394,41**	**6,41**	**2,59**	**2,59**	**1,88**	**13,47**	**26,94**	**Good**	**1,41**