Beyond Philology No. 16/4, 2019
ISSN 1732-1220, eISSN 2451-1498

https://doi.org/10.26881/bp.2019.4.03

# Method of measuring the effort related to post-editing machine translated outputs produced in the English>Polish language pair by Google, Microsoft and DeepL MT engines: A pilot study

MACIEJ KUR

*Received 02.09.2018,*
*received in revised form 15.07.2019,*
*accepted 04.09.2019.*

**Abstract**

This article presents the methodology and results of a pilot study concerning the impact of three popular and widely accessible machine translation engines (developed by Google, Microsoft and DeepL companies) on the pace of post-editing work and on the general effort related to post-editing of raw MT outputs. Fourteen volunteers were asked to translate and post-edit two source texts of similar characters and levels of complexity. The results of their work were collected and compared to develop a set of quantitative and qualitative data, which was later used to make assumptions related to the general rate of post-editing work and the quality of the post-edited sentences produced by the subjects. The aim of the pilot study described below was to determine whether the applied method can be successfully used in more profound studies on the quality and impact of machine translation in the English->Polish language pair and on the potential of MT solutions on the Polish translation market.

**Keywords**

machine translation, English->Polish language pair, post-editing, post-editing effort, pilot study, machine translation engines

## Nakład pracy podczas posteditingu. Badanie pilotażowe

**Abstrakt**

Niniejszy artykuł zawiera opis metodologii i wyników badania pilotażowego dotyczącego wpływu silników tłumaczenia maszynowego na tempo i nakład pracy związanej z postedycją. Czternaścioro ochotników dokonało tłumaczenia i postedycji dwóch tekstów źródłowych o podobnym charakterze i stopniu skomplikowania. Uzyskane wyniki zebrano i porównano, a na podstawie stworzonego w ten sposób zbioru danych ilościowych i jakościowych wyciągnięto ogólne wnioski dotyczące tempa i jakości pracy postedytora. Celem opisanego poniżej badania pilotażowego było określenie, czy zastosowana w nim metoda może zostać z powodzeniem wykorzystana podczas dogłębniejszych badań nad jakością i wpływem przekładu maszynowego w parze językowej angielski>polski oraz potencjałem rozwiązań MT na polskim rynku tłumaczeniowym.

**Słowa kluczowe**

przekład maszynowy, para językowa angielski>polski, postedycja, wysiłek postedycyjny, badanie pilotażowe, silniki tłumaczenia maszynowego

## 1. Introduction

Recent years have brought us considerable advances in the area of machine translation technology (MT) used to automatically translate source text materials into multiple target languages without human interference (cf. e.g. Bojar et al. 2016). Theoretically anticipated since the beginning of the 21st century, the neural machine translation systems, first introduced by Google

in October 2016 and later developed by other companies participating in the market, have caused substantial improvement in the quality of MT output (Bengio et al. 2003, Bentivogli et al. 2016, Wu et al. 2016). Consequently, MT engines have become more and more popular and widespread across the entire translation industry.[1] Nevertheless, the quality provided by even the most modern MT systems still requires human intervention in the form of post-editing, understood as a "correction of machine translation output by human linguists/editors" (Fiederer and O'Brien 2009, Allen 2003, Hutchins and Somers 1992).

Simultaneously, numerous scholars interested in MT technology have been conducting research aimed at the establishment of coherent and unified methods of MT output quality assessment (Bojar et al. 2016). Apart from traditional human-based methods that, for instance, involve the evaluations of the Fluency, Adequacy and Comprehension of machine-produced translations (Han et al. 2017), several Automatic Evaluation Metrics, such as BLEU or METEOR, have been developed to enable a reliable comparison of various MT engines and their efficiency with minimum human effort (Koehn 2010). As the overall quality of MT output depends on numerous factors, such as the source text type, language pair and target language (Hutchins and Somers 1992), the aforementioned task is a difficult one.

This article and the pilot study it describes constitute a preliminary step for a broader research project aimed at determining the level of MT output quality in the English->Polish language pair and the possibilities of applying popular MT systems in the Polish translation market. Such a task requires the establishment of a reliable, reproducible and cost-efficient study framework, which could be used to measure the efficiency of various solutions. The framework proposed in this article is based on several studies conducted in the past (Avramidis 2017, Graham et al. 2017, Han et al. 2017, Fiederer and O'Brien 2009, Callison-Burch et al. 2007, Snover et al. 2006) and is

[1] Cf. e.g. at <https://www.grandviewresearch.com/industry-analysis/machine-translation-market>, accessed 16.07.2018.

adjusted to the project's specific needs. The pilot study described below was designed in such a way to determine whether the proposed method can be efficiently used to obtain reliable results related to post-editing time and effort. It was conducted in May and June 2018 at the University of Gdańsk, Poland.

## 2. Methodology

### 2.1. Description

The pilot study was conducted on a test set composed of two source documents (A and B) containing 459 words each in 19 (Document A) and 20 (Document B) individual sentences. Both documents were presented to a group of 14 MA students in the Department of Translation Studies, who constituted a group of subjects. The subjects were divided into three groups (on the basis of the MT engine used to produce the output for post-editing – Google's GNMT, Microsoft's MNMT and DeepL Translator) and asked to perform a two-stage task based on translation and post-editing of the provided source material.

During the first stage of the task, the subjects were given 40 minutes to translate the contents of Document A into Polish and to produce target texts of the highest possible quality. During the second stage of the task, the subjects were given 40 minutes to post-edit the machine translated contents of Document B, again producing target texts of the highest possible quality. During both stages of the task, the number of parameters (total edit time, understood as time spent on editing target segments in the CAT software, words typed per hour and characters typed per hour, number of keystrokes and mouse clicks and numbers of "Delete" and "Backspace" key uses) were measured and recorded.

After the completion of the task, the data collected during both stages was gathered to enable the performance of quantitative and qualitative analyses of all parameters. Individual results obtained by the subjects during the first and second stages of the task were added and divided by the number of

participants assigned to the same group. This way, the average total edit time, words per hour, characters per hour, keystrokes, mouse clicks, backspace use and delete use parameters were obtained per each engine and each stage. These average values were later used to calculate the difference between individual engines and the difference between the parameters obtained during the translation stage and the post-editing stage.

The translations provided by each subject were collected and stored in the form of a .tmx file and were sent together with raw MT outputs to an independent translation agency for quality assessment. The reviewer was instructed to look at all translations of each source sentence and to order them on the basis of their quality. The results were then compared and analysed, and a general ranking of quality provided by individual subjects and all MT engines was developed.

In the meantime, the Human-targeted Translation Error Rate (HTER)[2] score was calculated for each translated sentence contained in the .tmx files on the basis of the MT outputs produced by the corresponding engines to measure the number of editing steps that needed to be performed by the subjects to obtain post-edited versions of the target text.

## 2.2. Source material

The entire research project was based on texts related to the construction industry, with particular reference to technical specifications. Hence the contents of both documents used during the pilot study were randomly picked from the specification of materials and workmanship,[3] describing general methods recommended during extending ground floors, altering interiors and converting the lofts of buildings. As the overall quality of the MT outputs relies heavily on the type of the processed source texts, the source material selected for such a study needs to be:

[2] Specia and Farzindar (2010); Snover et al. (2006).

[3] Publicly available at <http://studylib.net/doc/18186005/specification-of-materials-and-workmanship-required-in>, accessed 16.07.2018.

- highly repetitive in structure, preferably developed in accordance with a generally accepted pattern;
- focused more on terminology than on the style and other linguistic features;
- widely available.

We believe that the technical specifications meet all these requirements, as they are very often similar to each other and differ in details, their form is frequently governed by legal documents such as building codes applicable in particular countries (repetitive structure), they are technical in nature, their primary aim is to provide information (terminology focus) and they are abundant and can be easily obtained in both studied languages either on the internet or at construction companies (availability).

Initially, the test set developed for the pilot study contained two documents with 10 sentences (229 words in total) each, picked at random and meeting the abovementioned specification. After the analysis of results provided by the first two subjects it appeared, however, that Document B was too short. The subjects finished their task before the time was up and returned to the finished translations to review and redo them, which introduced unintentional noise into the obtained results. Therefore, the results obtained by subject 1 and subject 2 were excluded from the analysis of temporal parameters and the source texts volume was eventually doubled to reach 19 (Document A) and 20 sentences (Document B) (459 words in total) in each of the documents used during the pilot study.

The documents were cleared of any formatting and imported into separate CAT tools projects (one project for Document A and one project for Document B) with two separate and empty translation memories attached.

## 2.3. Subjects

Participants of the pilot study were recruited from the University of Gdańsk students who attended the MA course in the

Department of Translation Studies. The participation offer was entirely voluntary and aimed at students displaying sufficient motivation to take part in extracurricular activities. There were no preliminary requirements specified for the subjects, apart from being an MA student of Translation Studies. Interested students enrolled through an online form, where they specified the date and hour of their availability. In total, 14 subjects took part in the pilot study. The subjects were divided into three groups and each group worked on the contents processed by a different MT engine – Group A: Google Neural Machine Translation (GNMT – Subjects 1, 2, 3, 11, 12), Group B: Microsoft Neural Machine Translation (MNMT – Subjects 7, 8, 9, 10, 14) and Group C: DeepL Translator (DeepL – Subjects 4, 5, 6, 13).[4]

### 2.4. Session time

The pilot study sessions were designed to last 90 minutes. During that time, the subjects were provided with the instructions (5 minutes) and asked to perform both stages of the task (translation and post-editing – 40 minutes each). In between these two stages, a 5-minute break was organized for the subjects to rest and to allow for the collection of time-tracking and input data gathered during the first stage.

In total, five individual sessions were held between 21.05. 2018 and 9.06.2018, with one to seven subjects present simultaneously in the laboratory.[5]

### 2.5. Software and preparation

In order to unify the workspace and enable the collection of reliable time-tracking and input data, the subjects taking part in

[4] Unfortunately, the number of subjects recruited for the study could not be divided evenly, therefore the number of students working with one of the engines needed to be decreased.

[5] Because of this time span and due to the fact that the subjects were recruited from among the students of one university department, the possibility of communication taking place between various participants of the study in between sessions could not be completely ruled out.

the pilot study worked in MemoQ 8.2 and SDL Trados 2017 CAT tools. All subjects had some prior experience with MemoQ and SDL Trados software and both of these programs allow for time-tracking (either via an integrated feature in the case of MemoQ 8.2 or through a dedicated "Qualitivity" plugin in the case of SDL Trados 2017) and for the integration of machine translation engines. SDL Trados Studio 2017 was chosen mainly due to the lack of a proper plugin for DeepL integration in MemoQ 8.2 and it was used only by the subjects assigned to Group C.

Apart from time trackers, the computers used during the task were equipped with WhatPulse software,[6] which measured and registered the number of keystrokes and mouse clicks made by the subjects in their CAT tools. Apart from the overall amount, the software provided separate values corresponding to the use of the "Backspace" and "Delete" keys, which was especially helpful during the post-editing process.

Before each session, the workstations used by the subjects were prepared by creating new projects in CAT tools (for Document A and Document B separately) with empty translation memories and without term bases. In the case of projects with Document B imported, an applicable plugin for MT engine integration was enabled to allow for automatic translation of source segments with the use of the given engine directly in the CAT tool window. Similarly, the WhatPulse software was enabled in Windows OS and reset to delete all the data it may have accidentally registered before the beginning of each session.

### 2.6. Instructions and course of the session

Each session was supervised by an observer, who was responsible for the proper organization of the study and provided the task-related instructions. Before their appearance for the session at the laboratory, the subjects were not informed about the nature and focus of the pilot study and the information provided to them during the sessions was restricted to the minimum

[6] Available at <https://whatpulse.org/>, accessed 17.07.2018.

necessary. Before the commencement of work on the first stage of the task, the subjects were asked by the observer to:

- produce the translation of source sentences displayed on the screens of their workstations within the time limit of 40 minutes;
- focus on individual sentences and deliver as much high-quality target text as possible within the specified time frame;
- pay no attention to time limit, as the source document is purposefully too long to be translated completely on time;
- use all tools and techniques known and available to them apart from any MT engines and solutions;
- confirm every segment after the completion of each individual sentence.

After the break separating the first two stages of the task, the subjects were given identical instructions, with "Produce the translation" changed to "Post-edit". If any of the subjects was not familiar with the term, an oral explanation was given by the observer.

During the subjects' work, the observer's role was minimal. During the break and after the completion of the second stage, the observer collected all the time-tracking and input data gathered during the study and reset the CAT tools projects and the WhatPulse software.

## 2.7. Constraints

The pilot study described above was meant to test the proposed measurement method and to help to establish a sound and possibly reliable basis for more profound research work on MT technology in the English->Polish language pair. Due to its character, the study was subjected to various financial, temporal and organizational constraints that indirectly influenced the adopted methodology and obtained results. These constraints included the following factors, which need to be highlighted:

- The number of subjects was relatively low – among all students of the Translation Studies Department only 14 were willing to take part in the extracurricular study.
- Subjects were students with a low level of professional experience – the quality of the produced outputs does not mirror the quality of translation required from professional translators active on the translation market.
- Subjects could communicate with each other in between the sessions – due to the voluntary character of the study, the organization of a single simultaneous session for all subjects was impossible.
- The quality of the outputs produced by the subjects and MT engines during the first stage of the task was not compared with the quality produced during the second stage – the budget allocated for qualitative analysis made it possible to perform such an analysis on only one set of outputs.
- The comparison of the impact exerted by individual engines depended heavily on the skills of the given subject – due to temporal restrictions, each of the students worked on MT outputs produced by only one engine, which made it impossible to compare the results obtained by the same students with the use of various engines.

The problems listed above need to be taken into consideration during the interpretation of the results obtained during the study and the elimination of such problems is essential for any future research based on the proposed method. More details about potential areas for improvement are described in Section 5: Conclusions.

## 2.8. Results

## 3.1. Quantitative analysis

### 3.1.1. Stage 1

During the first stage of the task, the subjects were asked to deliver high-quality translations of the 20 sentences included in Document A in the time of 40 minutes. The time of their work

and some input methods (number of mouse clicks and key-strokes) were collected by the software installed and enabled at their workstations.

The total edit time required by the subjects to perform this stage varied from 33:53[7] to 44:43.[8] The average value of total edit time parameters obtained by all subjects was 39:49. As far as the Words per Hour parameter is concerned, the values obtained by the subjects varied from 207.73 to 602.97, while the value of the Characters per Hour parameter varied from 1572.64 to 3727.08. The average values for both these parameters were equal to 372.8 words per hour and 2442.83 characters per hour.

**Table 1**
Values of total edit time parameters measured
for each subject during Stage 1

| Subject No. | Total Edit Time |
|---|---|
| 3 | 00:39:40 |
| 4 | 00:40:29 |
| 5 | 00:40:18 |
| 6 | 00:41:26 |
| 7 | 00:35:51 |
| 8 | 00:39:43 |
| 9 | 00:39:09 |
| 10 | 00:44:43 |
| 11 | 00:33:53 |
| 12 | 00:40:09 |
| 13 | 00:42:37 |
| 14 | 00:39:45 |
| AVG | 00:39:49 |

[7] All temporal values are given in mm:hh format.

[8] In the case of subjects 4, 5, 6, 10, 12 and 13, the overall time of stage 1 completion exceeded the specified limit of 40 minutes, as some of the subjects refused to finish their work mid-sentence. The observer did not intervene in such cases.

**Table 2**
Values of words per hour and characters per hour parameters measured for each Subject during Stage 1

| Subject No. | Words per hour | Characters per hour |
|---|---|---|
| 3 | 461.25 | 3307.38 |
| 4 | 406.09 | 2042.32 |
| 5 | 602.97 | 3078.9 |
| 6 | 298.31 | 1572.64 |
| 7 | 312.95 | 2409.86 |
| 8 | 264.35 | 2054.37 |
| 9 | 444.43 | 3242.8 |
| 10 | 226.73 | 1770.94 |
| 11 | 517.01 | 3727.08 |
| 12 | 207.73 | 1579.64 |
| 13 | 464.61 | 2368.09 |
| 14 | 267.16 | 2159.89 |
| AVG | 372.80 | 2442.83 |

**Table 3**
Numbers of keystrokes, mouse clicks, "Backspace" and "Delete" keys uses measured for each subject during stage 1

| Subject No. | Keystrokes | Mouse Clicks | Backspace | Delete |
|---|---|---|---|---|
| 3 | 2857 | 151 | N/A | N/A |
| 4 | 2221 | 163 | N/A | N/A |
| 5 | 4475 | 193 | N/A | N/A |
| 6 | 2808 | 200 | 547 | 37 |
| 7 | 2544 | 156 | 237 | 0 |
| 8 | 1928 | 145 | 280 | 2 |
| 9 | 2947 | 146 | 307 | 8 |
| 10 | 1892 | 110 | 121 | 0 |
| 11 | 3366 | 119 | 387 | 10 |
| 12 | 1330 | 141 | 66 | 0 |
| 13 | 2629 | 218 | 67 | 4 |
| 14 | 2189 | 142 | 192 | 0 |
| AVG | 2598.83 | 157.00 | 244.89 | 6.78 |

The number of keystrokes used by the subjects in their CAT tools to complete stage 1 of the task varied from 1330 to 4475, with the average value equal to 2598.83, while the number of mouse clicks varied from 110 to 218, with the average value equal to 157. "Backspace" key was used between 66 and 54 times, while the "Delete" key was used between 0 and 37 times[9] (average values: 244.89 and 6.78).

### 2.8.1. Stage 2

During the second stage of the task, the subjects were asked to post-edit the machine translated contents of Document B. The parameters of their work were measured in a similar fashion as in the case of the first stage of the task.

Total Edit Time of post-edited segments varied from 17:52 to 41:35, with an average value equal to 33:18.

The number of words per hour typed by the subjects during the post-editing varied from 456.08 to 1538.06, while the number of characters ranged from 3467.07 to 7982.46, with average values equal to 817.6 words per hour and 5410.01 characters per hour.

As far as the input data collected during the second stage of the task is concerned, the number of keystrokes used during post-editing varied from 377 to 2206, with an average value equal to 1179.92, number of mouse clicks varied from 126 to 307, with an average value equal to 204.58, the "Backspace" key was used between 16 and 412 times, with an average value equal to 158.67, while the "Delete" key was used between 0 and 107 times, with an average value equal to 24.75.

[9] Due to a resetting mistake, the values corresponding to the use of "Backspace" and "Delete" keys by subjects 3, 4 and 5 were not properly registered and are therefore excluded from the described analysis.

**Table 4**
Values of total edit time parameters
measured for each Subject during Stage 2

| SUBJECT No. | Engine Used | Total Edit Time |
|---|---|---|
| 3 | GNMT | 00:30:12 |
| 4 | DeepL | 00:36:14 |
| 5 | DeepL | 00:34:48 |
| 6 | DeepL | 00:34:04 |
| 7 | MNMT | 00:41:35 |
| 8 | MNMT | 00:37:02 |
| 9 | MNMT | 00:35:00 |
| 10 | MNMT | 00:40:39 |
| 11 | GNMT | 00:33:50 |
| 12 | GNMT | 00:38:43 |
| 13 | DeepL | 00:17:52 |
| 14 | MNMT | 00:19:36 |
| AVERAGE | | 00:33:18 |

**Table 5**
Values of words per hour and characters per hour
parameters measured for each subject during stage 2

| Subject No. | Engine Used | Words per hour | Characters per hour |
|---|---|---|---|
| 3 | GNMT | 844.25 | 6350.73 |
| 4 | DeepL | 758.41 | 3936.15 |
| 5 | DeepL | 789.65 | 4098.27 |
| 6 | DeepL | 806.65 | 4186.49 |
| 7 | MNMT | 614.58 | 4613.68 |
| 8 | MNMT | 674.09 | 5097.78 |
| 9 | MNMT | 695.91 | 5315.31 |
| 10 | MNMT | 456.08 | 3467.07 |
| 11 | GNMT | 734.22 | 5542.14 |
| 12 | GNMT | 635.26 | 4838.84 |
| 13 | DeepL | 1538.06 | 7982.46 |
| 14 | MNMT | 1264.07 | 9491.2 |
| AVERAGE | | 817.60 | 5410.01 |

**Table 6**
Numbers of keystrokes, mouse clicks, "Backspace" and "Delete" keys uses measured for each Subject during Stage 2

| Subject No. | Engine Used | Key-strokes | Mouse Clicks | Back-space | Delete |
|---|---|---|---|---|---|
| 3 | GNMT | 1096 | 224 | 115 | 107 |
| 4 | DeepL | 645 | 176 | 55 | 0 |
| 5 | DeepL | 1031 | 297 | 112 | 18 |
| 6 | DeepL | 1076 | 307 | 253 | 0 |
| 7 | MNMT | 2206 | 136 | 287 | 0 |
| 8 | MNMT | 1202 | 238 | 203 | 0 |
| 9 | MNMT | 1678 | 133 | 177 | 106 |
| 10 | MNMT | 1908 | 284 | 136 | 0 |
| 11 | GNMT | 1704 | 190 | 412 | 58 |
| 12 | GNMT | 770 | 188 | 89 | 0 |
| 13 | DeepL | 377 | 126 | 16 | 8 |
| 14 | MNMT | 466 | 156 | 49 | 0 |
| AVERAGE | | 1089.36 | 191.00 | 149.14 | 21.21 |

In the case of the Total Edit Time parameter, only Subject 7 needed more time to complete post-editing stage of the task in comparison with the time required for translation without the aid of an MT engine (15.99% difference); all other subjects worked more quickly when post-editing the machine translated sentences than during translation without MT support, with differences varying from -0.15% to -58.08%. The average difference calculated on the basis of all of the collected results was equal to -15.73%.

The number of words typed per hour during the post-editing stage of the task increased between 30.96% and 373.15% in comparison to stage 1, giving an average value of 136.03%, while the number of characters typed per hour increased between 33.11% and 339.43%, with the average value equal to 134.57%.

**Table 7**
Difference in the values of total edit time parameters measured for each subject during stage 1 and 2

| Subject No. | Engine Used | Difference - time |
|---|---|---|
| 3 | GNMT | -23.87% |
| 4 | DeepL | -10.50% |
| 5 | DeepL | -13.65% |
| 6 | DeepL | -17.78% |
| 7 | MNMT | 15.99% |
| 8 | MNMT | -6.76% |
| 9 | MNMT | -10.60% |
| 10 | MNMT | -9.09% |
| 11 | GNMT | -0.15% |
| 12 | GNMT | -3.57% |
| 13 | DeepL | -58.08% |
| 14 | MNMT | -50.69% |
| AVERAGE | | -15.73% |

**Table 8**
Difference in the values of words per hour and characters per hour parameters measured for each subject during Stage 1 and 2

| Subject No. | Engine Used | Difference – words | Difference – characters |
|---|---|---|---|
| 3 | GNMT | 83.04% | 92.02% |
| 4 | DeepL | 86.76% | 92.73% |
| 5 | DeepL | 30.96% | 33.11% |
| 6 | DeepL | 170.41% | 166.21% |
| 7 | MNMT | 96.38% | 91.45% |
| 8 | MNMT | 155.00% | 148.14% |
| 9 | MNMT | 56.58% | 63.91% |
| 10 | MNMT | 101.16% | 95.78% |
| 11 | GNMT | 42.01% | 48.70% |
| 12 | GNMT | 205.81% | 206.33% |
| 13 | DeepL | 231.04% | 237.08% |
| 14 | MNMT | 373.15% | 339.43% |
| AVERAGE | | 136.03% | 134.57% |

A comparison of the input parameters obtained during the performance of the first two stages of the task revealed that most of the subjects used fewer keystrokes and more mouse clicks during post-editing than during human translation (average values equal to respectively -51.69% and 35.41%). When compared to human translation, the "Backspace" key was used less frequently (-22.15% on average) and the "Delete" key was used more frequently (178.33% on average) in the case of post-editing.

**Table 9**
Difference in the numbers of keystrokes, mouse clicks, "Backspace" and "Delete" keys uses measured for each subject during stage 1 and 2

| Subject No. | Engine Used | Difference – Keystrokes | Difference – Mouse Clicks | Difference – Backspace | Difference – Delete |
|---|---|---|---|---|---|
| 3 | GNMT | -61.64% | 48.34% | N/A | N/A |
| 4 | DeepL | -70.96% | 7.98% | N/A | N/A |
| 5 | DeepL | -76.96% | 53.89% | N/A | N/A |
| 6 | DeepL | -61.68% | 53.50% | -53.75% | -100.00% |
| 7 | MNMT | -13.29% | -12.82% | 21.10% | 0.00% |
| 8 | MNMT | -37.66% | 64.14% | -27.50% | -100.00% |
| 9 | MNMT | -43.06% | -8.90% | -42.35% | 1225.00% |
| 10 | MNMT | 0.85% | 158.18% | 12.40% | 0.00% |
| 11 | GNMT | -49.38% | 59.66% | 6.46% | 480.00% |
| 12 | GNMT | -42.11% | 33.33% | 34.85% | 0.00% |
| 13 | DeepL | -85.66% | -42.20% | -76.12% | 100.00% |
| 14 | MNMT | -78.71% | 9.86% | -74.48% | 0.00% |
| AVERAGE | | -51.69% | 35.41% | -22.15% | 178.33% |

### 2.8.2. HTER scoring

The post-edited sentences provided by the subjects were used to calculate HTER scores for each sentence and each MT engine

used during stage 2 of the study. HTER is a metric that can be efficiently used to measure the effort required to post-edit MT outputs and, in these terms, to evaluate the efficiency and usefulness of MT engines (Specia and Farzindar 2010; Snover et al. 2006).

Individual HTER scores were calculated with the use of the following formula:

HTER= # of editing steps# of reference words

where editing steps included all insertions, deletions, substitutions and shifts of word sequences used by particular subjects during the post-editing stage of the task to produce target sentences, treated during the analysis as "reference" translations, giving the results between "0" (no editing steps) to "1" (the MT output changed entirely).

The HTER scores were calculated for each source sentence in relation to each target-reference sentence provided by the subjects. The results were then averaged per sentence and per engine to allow for a comparative analysis of effort and impact of MT outputs provided by all three MT engines used during the study.

HTER scores obtained during the performance of stage 2 by individual subjects varied from 0.11 to 0.62, giving an average HTER score of 0.30.

The subjects who post-edited outputs provided by GNMT achieved average HTER scores in the range between 0.12 (Sentence 20) and 0.58 (Sentence 14), with an overall average score equal to 0.29. When working on outputs provided by MNMT, HTER scores obtained by the subjects varied from 0.05 (Sentence 11) to 0.69 (Sentences 8 and 17), with an overall average score equal to 0.43. The DeepL engine allowed for the obtainment of HTER scores in the range between 0.06 (Sentences 11 and 19) and 0.32, with an overall average score equal to 0.17.

**Table 10**
Average HTER scores calculated for each of the subjects

| Avg HTER score per Subject | |
|---|---|
| Subject no. | Avg HTER score |
| 1 | 0.16 |
| 2 | 0.25 |
| 3 | 0.34 |
| 4 | 0.18 |
| 5 | 0.22 |
| 6 | 0.16 |
| 7 | 0.47 |
| 8 | 0.38 |
| 9 | 0.43 |
| 10 | 0.62 |
| 11 | 0.36 |
| 12 | 0.23 |
| 13 | 0.11 |
| 14 | 0.34 |
| TOTAL AVG | 0.30 |

**Table 11**
Average HTER scores calculated for each sentence and MT engine

| Average HTER scores per sentence and per engine | | | |
|---|---|---|---|
| Sentence no. | GNMT | MNMT | DeepL MT |
| 1 | 0.25 | 0.37 | 0.15 |
| 2 | 0.21 | 0.44 | 0.18 |
| 3 | 0.15 | 0.23 | 0.20 |
| 4 | 0.31 | 0.33 | 0.17 |
| 5 | 0.29 | 0.49 | 0.27 |
| 6 | 0.18 | 0.62 | 0.13 |
| 7 | 0.30 | 0.61 | 0.18 |
| 8 | 0.33 | 0.69 | 0.09 |
| 9 | 0.16 | 0.37 | 0.10 |
| 10 | 0.54 | 0.42 | 0.28 |

| | | | |
|---|---|---|---|
| 11 | 0.37 | 0.05 | 0.06 |
| 12 | 0.38 | 0.24 | 0.08 |
| 13 | 0.29 | 0.40 | 0.20 |
| 14 | 0.58 | 0.53 | 0.26 |
| 15 | 0.16 | 0.49 | 0.21 |
| 16 | 0.37 | 0.53 | 0.32 |
| 17 | 0.45 | 0.69 | 0.18 |
| 18 | 0.16 | 0.23 | 0.13 |
| 19 | 0.15 | 0.43 | 0.06 |
| 20 | 0.12 | 0.61 | 0.16 |
| TOTAL AVG | 0.29 | 0.43 | 0.17 |

## 2.9. Qualitative analysis

All of the sentences produced by the subjects during stage 2 of the task were collected from .tmx files, placed in a single .xls file, anonymised and randomized to change their order and prevent any accidental pattern recognition. The .xls file was then sent to an independent translation agency with a request to evaluate the quality of target sentences and to put them in order in terms of their quality.

The evaluator was given a set of instructions to follow and was required to take into consideration the accuracy, fluency and style of the provided sentences.[10] Apart from the target sentences produced by the subjects, the .xls file contained also the un-post-edited MT outputs. No information about the subject of the study and machine translated outputs used as a basis for post-editing was revealed to the evaluator.

The evaluated sentences were given points from "1" (the highest quality) to "17"[11] (the lowest quality). The potential ideal

[10] Understood as: "accuracy" – the translation should contain the same information as the source text; "fluency" – the translation should be easily understandable for the reader of the target text; "style" – the translation should be adjusted to the character and aim of the source text (cf. White, 1994).

[11] As the collected outputs included 14 sentences post-edited by the Subjects and 3 machine translated sentences for each source segment.

output would therefore obtain a score of 20 points, and the potential worst output would obtain a score of 304 points.[12] If the evaluator decided that two or more translations represented similar quality, an equal score could be given to them.

The scores assigned to particular sentences were later collected, grouped and averaged per subject and per engine. The scores given to raw MT outputs were then compared with the results obtained by particular subjects.

**Table 12**

Sums and average numbers of quality points given to each of the subjects and MT engines

| Subject no. (MT engine used) | Sum of quality points (# of produced sentences) | Avg quality points |
|---|---|---|
| 1 (GNMT) | 64 (9) | 8.00 |
| 2 (GNMT) | 61 (10) | 6.10 |
| 3 (GNMT) | 90 (20) | 4.50 |
| 4 (DeepL) | 122 (20) | 6.10 |
| 5 (DeepL) | 127 (20) | 6.35 |
| 6 (DeepL) | 114 (20) | 5.70 |
| 7 (MNMT) | 173 (19) | 9.11 |
| 8 (MNMT) | 164 (20) | 8.20 |
| 9 (MNMT) | 234 (20) | 11.70 |
| 10 (MNMT) | 42 (9) | 5.25 |
| 11 (GNMT) | 120 (20) | 6.00 |
| 12 (GNMT) | 112 (18) | 6.22 |
| 13 (DeepL) | 154 (20) | 7.70 |
| 14 (MNMT) | 225 (19) | 11.84 |
| Scores given to raw MT outputs | | |
| GNMT | 212 (20) | 10.6 |
| MNMT | 212 (20) | 10.6 |
| DeepL | 124 (20) | 6.2 |

[12] The overall range of points was lowered for the sentences that were not post-edited by all Subjects due to the time restriction.

The sum of points given to individual subjects varied from 90 to 234,[13] with 212 points given to raw outputs provided by GNMT and MNMT engines and 124 points given to raw outputs provided by DeepL engine. The average numbers of points given to the subjects varied from 4.50 to 11.84 per sentence, with 10.6 given to GNMT and MNMT engines and 6.2 given to DeepL engine. The following table presents a summary of total and average numbers of quality points assigned to individual subjects and MT engines.

## 3. Summary

The primary aim of the task was to determine whether the method described above could be used to efficiently measure the time and effort required during translation and post-editing of texts of similar volumes and levels of complexity. The measurement method was based on the comparison of temporal results obtained during stage 1 and stage 2 of the task and on the comparison of HTER scores calculated for each post-edited sentence produced by the subjects. The following section presents the overall summary of results obtained during the study.

In general, when time of work is considered, the subjects worked 15.73% faster during the post-editing of raw MT outputs than when translating texts from scratch. During stage 2 of the task they also entered 136.03% words and 134.57% characters more in target segments within a unit of time and they needed 51.69% keystrokes less in comparison with stage 1. Simultaneously, during the post-editing stage of the task the subjects needed 35.41% more mouse clicks to perform their work than they did during the translation stage. Considering the character of the post-editing work and the number of editing steps that needed to be introduced during the process in various parts of sentences provided by MT engines, such results could be anticipated, as numerous words and fragments of sentences

[13] Some subjects did not manage to post-edit all sentences given to them during stage 2 of the task. The total numbers of sentences post-edited by particular subjects are given in parentheses.

were already placed in the target segments and the subjects needed less time and keystrokes to edit them, but more mouse clicks to actually place their cursors in places that required editing.

Considering the impact of individual MT engines used during the study on the time of subjects' work, the seemingly best results were obtained with the use DeepL engine (25% shorter editing time), followed by MNMT (12.23% shorter editing time) and GNMT (6.14% shorter editing time).

**Table 13**
Differences in the values of key parameters measured during the study for each MT engine

| Average temporal and input parameters per engine | | | | | |
|---|---|---|---|---|---|
| Engine | Difference – time | Difference – words | Difference – characters | Difference – keystrokes | Difference – mouse clicks |
| GNMT | -6.14% | 76.60% | 81.11% | -58.92% | 29.05% |
| MSMT | -12.23% | 156.45% | 147.74% | -34.37% | 42.09% |
| DeepL | -25.00% | 129.79% | 132.28% | -73.82% | 18.29% |

The results obtained for the use of the "Backspace" and "Delete" keys were inconsistent, ranging from -76.12% to 34.85% in the case of the former and -100% to 1225% in the case of the latter, leading to the conclusion that the use of these keys is an individual matter.

As far as the HTER scores are considered, the overall average score obtained by the Subjects (0.30) could be perceived as comparable with the results achieved during some previous studies (Snover 2006). As HTER scoring depends greatly on the post-editorial skills, it can be presumed that it would be higher in the case of the same task given to more skilled subjects. There was a visible tendency displayed by some of the subjects, who did not introduce many changes in sentences they considered "acceptable". As the subjects were not professional translators or

specialists in the area of construction industry, some post-edited sentences copied the mistakes included in raw MT output. As such mistakes were not rectified, the HTER scores could be lowered.

The lowest HTER scores were achieved by the subjects working on outputs provided by DeepL engine (average HTER score: 0.17), followed by GNMT (0.29) and MNMT (0.43). To some extent, a relation between HTER score and Total Edit Time parameter can be observed, as subjects working on outputs provided by DeepL engine were both fastest and obtained the lowest HTER scores on average. The same could be said about the results obtained by Subject 13 (the highest increase in time – 58.08% and the lowest HTER score – 0.11). However, when considering the results obtained with the use of GNMT and MNMT, there seems to be no direct relation between pace of work (6.14% and 12.23% respectively) and HTER scores (0.29 and 0.43 respectively). Similarly, the HTER results obtained by Subjects 14 and 3, who followed Subject 13 in terms of post-editing speed (50.69% and 23.87% faster than when translating from scratch) were higher than for instance the HTER scores obtained by Subjects 4 and 12, who were much slower in post-editing than Subject 13 (10.50% and 3.57% respectively). Figure 1 depicts the aforementioned relation.

The difference in the quality of post-edited sentences provided by the subjects was determined on the basis of a ranking developed by an independent translation agency. The lowest average number of points was obtained by the subjects working on the outputs produced by the GNMT engine (5.53), followed by DeepL (6.46) and MNMT (9.69), with the lowest score achieved by Subject 3 (GNMT – 4.50) and the highest score achieved by Subject 14 (MNMT – 11.84). Table 14 contains the summary of temporal, HTER and qualitative results obtained by each of the subjects.

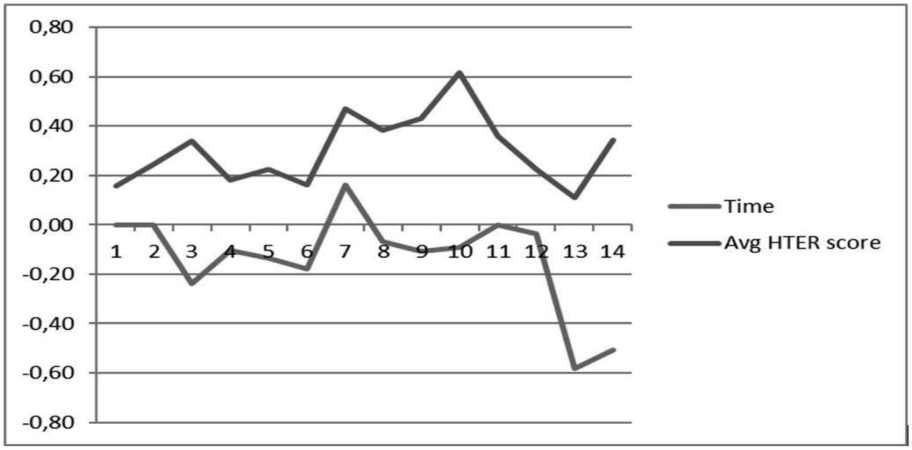


**Figure 1**
Relation between total edit time parameter and average HTER scores

**Table 14**
Comparison of total edit time parameters, average HTER scores and average number of quality points

| Comparison of temporal, HTER and qualitative parameters per Subject | | | |
|---|---|---|---|
| Subject no. | Editing time difference | Avg HTER score | QA AVG |
| 1 | n/a | 0.16 | 8.00 |
| 2 | n/a | 0.25 | 6.10 |
| 3 | -23.87% | 0.34 | 4.50 |
| 4 | -10.50% | 0.18 | 6.10 |
| 5 | -13.65% | 0.22 | 6.35 |
| 6 | -17.78% | 0.16 | 5.70 |
| 7 | 15.99% | 0.47 | 9.11 |
| 8 | -6.76% | 0.38 | 8.20 |
| 9 | -10.60% | 0.43 | 11.70 |
| 10 | -9.09% | 0.62 | 5.25 |
| 11 | -0.15% | 0.36 | 6.00 |
| 12 | -3.57% | 0.23 | 6.22 |
| 13 | -58.08% | 0.11 | 7.70 |
| 14 | -50.69% | 0.34 | 11.84 |
| TOTAL AVG | -15.73% | 0.30 | 7.34 |

The quality of raw MT outputs was also evaluated and the individual engines achieved average scores of 10.6 (GNMT and MNMT[14]) and 6.2 (DeepL). In most cases, the raw MT outputs were evaluated lower than post-edited sentences, with most notable exception of 3 sentences translated by DeepL engine, which were considered to be of the highest quality (score = 1) among all other provided translations. In general however, the outputs provided by MT engines were evaluated lower than the outputs provided by most human subjects.

Figure 2 depicts the relation between all 3 key parameters measured during the pilot study.

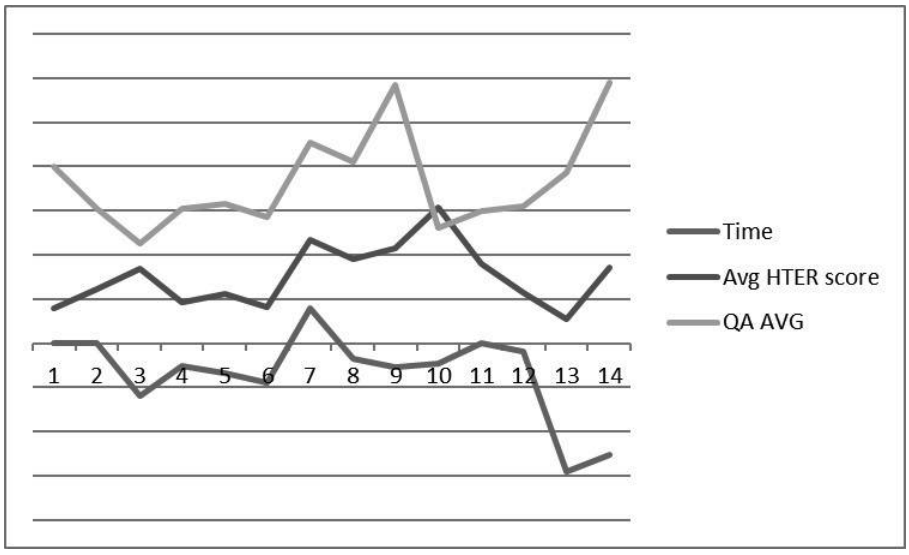


**Figure 2**
Relation between total edit time parameter, average HTER scores and average number of quality points

[14] The scores given to most individual sentences translated by GNMT and MNMT varied, the equal overall average results seem to be a mere coincidence.

In the case of some subjects (most notably Subjects 6, 7, 8, 13 and 14) there seems to be a recognizable pattern of relation between the parameters. However, the determination of a general dependence would require some further research.

## 4. Conclusions

The analysis of the practical implementation of the measurement method proposed above and the results obtained with its use revealed several drawbacks that should be considered before applying the aforementioned methodology in more profound studies on the impact of MT solutions on post-editing effort and quality of the final product in the English->Polish language pair. Some of these drawbacks resulted from the constraints that influenced the pilot study[15], others were caused by the lack of experience in organization of similar research. Nevertheless, the experiences gained during the pilot study allowed for the identification of potential areas for improvement that should be implemented in the future studies in order to obtain more reliable results. The following list presents these areas:

- Subjects should be recruited from among professional translators with more experience in translation and post-editing than students, to obtain outputs of higher quality.
- Groups should be composed of as many subjects as possible, to obtain higher reliability of the average scores and results.
- Sessions should take place simultaneously, to make it impossible for the subjects to communicate amongst themselves.
- Source texts should be long enough, to eliminate the risk of subjects finishing their tasks before time.
- Subjects should be given a clear signal to stop working, to avoid unnecessary noise in temporal data.
- Much attention should be paid to the preparatory stage taking place before sessions, to avoid any resetting mistakes that could negatively influence the consistency of results.
- Each of the subjects should be given an opportunity to produce post-edited target segments with the support of each MT engine

[15] Cf. Section 2.7: Constraints.

analysed during the study, to allow for direct comparison of results independent of individual skills of the subjects.
- The quality of outputs produced during stage 1 of the task should be evaluated in a similar manner as the quality of post-edited segments, to allow for comparative analysis of both methods of target text production.

We believe that careful consideration of these areas and their implementation during future research work would improve the general usefulness and efficiency of the proposed methodology, which in its amended form could be successfully used to measure various parameters related to post-editing in the English->Polish language pair and to obtain reliable and repetitive results.

## 5. Further study

The pilot study described in this article was designed and conducted as an attempt to establish a reliable methodological basis for more detailed research on machine translation in the English->Polish language pair. The list of potential areas for improvement presented above will be used to develop the method further and to obtain more reliable, repetitive and standardized results and patterns.

Future research will involve the development of a larger corpus of technical and construction industry texts and translations, detection and categorization of the most common errors occurring in MT output materials, development and training of a Moses-based MT engine and performance of tests aimed at the determination of possible potential of regular expressions in automatic post-editing and improvement of MT output quality. The results of these efforts will be presented in future publications.

## References

Allen, Jeffrey (2003). "Post-editing". In: Harold Somers (ed.). *Computers and Translation: A Translator's Guide*. Amsterdam – Philadelphia: John Benjamins, 297-317.

Avramidis, Eleftherios (2017). "Comparative quality estimation for machine translation observations on machine learning and features". *The Prague Bulletin of Mathematical Linguistics* 108/1: 307-318.

Bengio, Yoshua, Rejean Ducharme, Pascal Vincent, Christian Jauvin (2003). "A neural probabilistic language model". *Journal of Machine Learning Research* 3: 1137-1155.

Bentivogli, Luisa, Arianna Bisazza, Mauro Cettolo, Marcello Federico (2016). "Neural versus phrase-based machine translation quality: A case study". In: Jian Su, Kevin Duh, Xavier Carreras (eds.). *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, USA: Association for Computational Linguistics, 257-267.

Bojar, Ondrej, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, Marcos Zampieri (2016). "Findings of the 2016 Conference on Machine Translation (WMT16)". In: Ondrej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, Marcos Zampieri (eds.). *Proceedings of the First Conference on Machine Translation*. Volume 2: *Shared Task Papers*. Berlin: Association for Computational Linguistics, 131-198.

Callison-Burch, Chris, Cameron Fordyce, Philipp Koehn, Christof Monz, Josh Schroeder (2007). "(Meta-) Evaluation of machine translation". In: Chris Callison-Burch, Philipp Koehn (eds.). *StatMT '07 Proceedings of the Second Workshop on Statistical Machine Translation*. Stroudsburg, USA: Association for Computational Linguistics, 136-158.

Fiederer, Rebecca, Sharon O'Brien (2009). "Quality and machine translation: A realistic objective?". *The Journal of Specialised Trans-*

*lation* 11: 52-74. Available at <https://www.jostrans.org/issue11/art_fiederer_obrien.pdf>. Accessed 12.07.2019.

Graham, Yvette, Quingsong Ma, Timothy Baldwin, Qun Liu, Carla Parra, Carolina Scarton, (2017). "Improving evaluation of document-level machine translation quality estimation". In: Mirella Lapata, Phil Blunsom, Alexander Koller (eds.). *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*: Volume 2: *Short Papers*. Valencia: Association for Computational Linguistics, 356-361.

Han, Lifeng, Derek F. Wong, Lidia S. Chao (2017). "Machine Translation Evaluation Resources and Methods: A Survey". Cornell University Library. Available at <https://arxiv.org/abs/1605.04515>. Accessed 12.07.2019.

Hutchins, John, Harold L. Somers (1992). *An Introduction to Machine Translation*. London: Academic Press.

Koehn, Philipp (2010). *Statistical Machine Translation*. Cambridge: Cambridge University Press.

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, John Makhoul (2006). "A study of translation edit rate with targeted human annotation". In: Laurie Gerber, Nizar Habash, Alon Lavie (eds.). *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*. Cambridge, USA: AMTA, 223-231.

Specia, Lucia, Atefeh Farzindar (2010). "Estimating machine translation post-editing effort with HTER". Conference Paper at AMTA 2010-workshop, Bringing MT to the User: MT Research and the Translation Industry. Denver, USA, 31.10–4.11.2010.

White, John, Theresa O'Connel, Francis O'Mara (1994). "The ARPA MT evaluation methodologies: Evolution, lessons, and future approaches". In: Muriel Vasconcellos( ed.). *Proceedings of the First Conference of the Association for Machine Translation in the Americas*. Columbia, USA: AMTA, 193-205.

Wu, Yonghui, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi (2016). "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation". Cornell University Library. Available at <https://arxiv.org/pdf/1609.08144.pdf>. Accessed 12.07.2019.

Maciej Kur
ORCID iD: 0000-0002-7344-5372
University of Gdańsk
Institute of English and American Studies
Wita Stwosza 51
80-308 Gdańsk
Poland
maciej.kur@ug.edu.pl