informatics

MDPI

Article

The Role of Machine Translation Quality Estimation in the Post-Editing Workflow

Hannah Béchara [1,*], Constantin Orăsan [2], Carla Parra Escartín [3], Marcos Zampieri [4] and William Lowe [1]

1 Hertie School Data Science Lab., 10117 Berlin, Germany; lowe@hertie-school.org
2 Centre of Translation Studies, University of Surrey, Guildford GU2 7XH, UK; c.orasan@surrey.ac.uk
3 RWS Language Weaver, Dublin, Ireland; cparraescartin@rws.com
4 Rochester Institute of Technology, Department of Computer Science, Rochester, NY 14623, USA; Marcos.Zampieri@rit.edu
* Correspondence: hjbechara@gmail.com or bechara@hertie-school.org

Abstract: As Machine Translation (MT) becomes increasingly ubiquitous, so does its use in professional translation workflows. However, its proliferation in the translation industry has brought about new challenges in the field of Post-Editing (PE). We are now faced with a need to find effective tools to assess the quality of MT systems to avoid underpayments and mistrust by professional translators. In this scenario, one promising field of study is MT Quality Estimation (MTQE), as this aims to determine the quality of an automatic translation and, indirectly, its degree of post-editing difficulty. However, its impact on the translation workflows and the translators' cognitive load is still to be fully explored. We report on the results of an impact study engaging professional translators in PE tasks using MTQE. To assess the translators' cognitive load we measure their productivity both in terms of time and effort (keystrokes) in three different scenarios: translating from scratch, post-editing without using MTQE, and post-editing using MTQE. Our results show that good MTQE information can improve post-editing efficiency and decrease the cognitive load on translators. This is especially true for cases with low MT quality.

Keywords: machine translation quality estimation; post-editing

Citation: Béchara, H.; Orăsan, C.; Parra Escartín, C.; Zampieri, M.; Lowe, W. The Role of Machine Translation Quality Estimation in the Post-Editing Workflow. *Informatics* **2021**, *8*, 61. https://doi.org/10.3390/informatics8030061

Academic Editor: Antony Bryant

Received: 15 July 2021
Accepted: 6 September 2021
Published: 14 September 2021

1. Introduction

According to recent user surveys [1,2] the use of Machine Translation (MT) has become more popular among professional translators in the last decade. As Machine Translation Post-Editing (MTPE) becomes more widely used in the translation industry, assessing the quality of the MT becomes a more pressing concern. A poor MT suggestion might end up taking the post-editor more time to assess and rewrite than it would to translate from scratch [3]. (For the purposes of this paper, we refer to post-editing as a task usually carried out by professional translators, and hence we use the term professional translators as an umbrella for both translators and post-editors.) Consistent low-quality MT can prompt professional translators to give up on post-editing and revert to translating from scratch. We therefore posit that a system that assesses the quality of the MT suggestions can help prevent frustration and cut down on the time it takes the post-editor to decide if a suggestion is worth post-editing. Machine Translation Quality Estimation (MTQE) can provide this assessment and help the post-editor by only proposing sentences which are good enough to be post-edited.

This paper investigates the effect of Machine Translation Quality Estimation (MTQE) in the MTPE pipeline in a professional setting. MTQE automatically estimates the quality of machine translation output without relying on a reference translation. Unlike usual metrics used to evaluate MT, MTQE does not require human intervention to provide a quality estimation. We design and present a user study that enlists the help of professional translators with post-editing of real-world data. We employ a traffic light system to present

translators with different categories of sentences and determine how effective MTQE is at improving the efficiency of the translation workflow. Four different professional translators are engaged in this task, and their productivity is measured when translating without an MT suggestion, post-editing without MTQE assessment, and post-editing with MTQE information. We find that good MTQE information can improve the post-editing efficiency. This paper builds on our preliminary published work [4] by carrying out a much more detailed analysis of the data.

The rest of this paper is organised as follows: In Section 2, we provide an overview of the previous research on this subject. Section 3 presents our approach to using MTQE to label the data. Section 4 describes the user study and provides the details of the traffic light system. Section 5 presents our results, analysis, and discussion around the findings. Finally, Section 6 sums up our research and presents possible future avenues of investigation.

2. Background

It has been shown in a number of studies that the use of MT by professional translators often leads to productivity gains which are, in turn, related to the quality of the MT segments provided to the translators [5,6]. To ensure that high quality segments are presented to the translator, MTQE or other forms of quality assessments have to be employed. The aim of MTQE is to provide an accurate assessment of a given translation without input from a reference translation or a human evaluator. To date, and as pointed out by Turchi et al. [7], "QE research has not been followed by conclusive results that demonstrate whether the use of quality labels can actually lead to noticeable productivity gains in the CAT framework". This suggests that how we use MTQE in the translation workflows is still an open question to be answered, despite previous work integrating MTQE into them.

Previous research into the use of MTQE in a professional setting has been carried out by researchers from translation studies and related fields. Most notably, Turchi et al. [7] ran a similar study to ours, and investigated whether the use of binary labels (green for a good MT suggestion and red for a bad one) can significantly improve the productivity of translators. The authors used MateCat [8], adapted to provide a single MT suggestion, and a red or green label. They chose a HTER [9] of 0.4 as the boundary between post-editing and translating from scratch. HTER ("Human-targeted Translation Error Rate") is a human-targeted edit distance measure. HTER is based on both the difficulty of the translation, and the discrepancies between the source and target sentences. A lower HTER is desirable for a higher quality translation. Therefore, all sentences with a predicted HTER score under 0.4 were labelled green, and any sentence with a HTER over 0.4 was labelled red. Their dataset was an English user manual in the IT domain, translated into Italian using the phrase-based SMT system Moses [10] and then post-edited. In total, their dataset consisted of 1389 segments, of which 542 were used to train the MTQE engine, and 847 were used for testing. In total, they gathered two instances of each segment, one for the scenario in which the translator was shown the estimated quality of the translation, and one in which the translator did not have this information for the MT output. While they observed a slight increase in productivity (1.5 s per word), they concluded that this increase is not statistically significant across the dataset. However, further investigation of their data showed a statistically significant percentage of gains for medium-length suggestions with HTER > 0.1. Our user study follows in the footsteps of this study, with several differences. We use the Fuzzy Match Score (FMS) instead of more traditional MT evaluation metrics as translators are more used to working with TM leveraging and fuzzy matches [11,12]. Fuzzy match scores approximately match strings. The fuzzy match score can range from 0 through 100, based on how close the hypothesis and reference sentences match. While several algorithms for FMS exist, we use the fuzzy value computed by Okapi Rainbow (http://okapi.opentag.com/, accessed on 9 September 2021). The comparison is based in 3-g at the character level. Furthermore, our experimental setup accounts for two "neutral" conditions, also shown to the translators as different colours: one in which they translate from scratch, and one in which they are shown MT, but not MTQE information, and are

asked to decide whether they post-edit or translate from scratch. Most significantly, our study attempts to look at the difference between the effects of good MTQE, versus that of mediocre or even inaccurate MTQE.

Also in 2015, Moorkens et al. [13] investigated the accuracy of human estimates of post-editing effort and the extent to which they mimic actual post-editing effort. They also researched how much the display of confidence scores (MTQE) influenced post-editing behaviour. The authors used two different groups of participants. The first consisted of six members of staff and researchers who were not students of translation or professional translators. The second group consisted of 33 translation students, both at the masters and undergraduate levels. The first group was presented with a set of 80 machine translated segments from two Wikipedia articles describing Paraguay and Bolivia. The sentences were translated into Portuguese using Microsoft Bing Translator (https://www.bing.com/translator, accessed on 9 September 2021). The experiment consisted of three phases. In the first phase, the translators were asked to classify the MT output according to the following scale:

1. Segment requires a complete re-translation;
2. Segment requires some post-editing; and
3. Segment requires little or no post-editing.

In the second phase, the same participants were asked to post-edit these segments. In order to avoid the participants remembering their ratings, they were given two weeks in between the evaluation phase and the post-editing phase of the experiments. The researchers then used the ratings collected in the first phase of this experiment to create a set of MTQE labels for the segments, and in the third phase, they presented the segments with these labels to the second group of participants (undergraduate and masters students). The second group were asked to post-edit the segments, using the MTQE labels. While their study only spans 80 segments, their findings suggest that "the presentation of post-editing effort indicators in the user interface appears not to impact on actual post-editing effort". Their conclusions contradict our findings, which we will expand on later in this paper.

Moorkens and Way [14] compared the use of translation memory (TM) to that of MT among translators. They presented 7 translators with 60 segments of English—German translations, extracted from the documentation of an open-source program called freeCAD (https://www.freecadweb.org/, accessed on 9 September 2021) and from the Wikipedia entry on the same topic. Their results show that low-quality MT matches are not useful to the translators in over 36% of cases. Furthermore, the translators described these suggestions as "irritating". In contrast, TM matches were always found to be useful. Moorkens and Way [14] conclude that their findings suggest that "MT confidence measures need to be developed as a matter of urgency, which can be used by post-editors to wrest control over what MT outputs they wish to see, and perhaps more importantly still, which ones should be withheld".

In their more recent work, Moorkens and O'Brien [15] attempt to determine the specific user interface needs for post-editors of MT. The authors conduct a survey of translators and report that 81% expressed the need for confidence scores for each target text segment from the MT engine. This result validates the impact of the study reported in this paper, as we precisely investigate the effect of showing MTQE to translators when undergoing post-editing tasks.

Teixeira and O'Brien [16] investigated the impact of MTQE on the post-editing effort of 20 English to Spanish professional translators post-editing four texts from the WMT13 news dataset (http://www.statmt.org/wmt13/, accessed on 9 September 2021). They used four types of scenarios: No MTQE, Accurate MTQE, Inaccurate MTQE, and Human Quality Estimation, which was calculated using the direct assessment method proposed by Graham et al. [17]. Their goal was to determine the impact of the different modes of MTQE on the time spent (temporal effort), the number of keystrokes (physical effort), and the gaze behaviour (cognitive effort). Their results showed no significant differences in terms of cognitive effort. In the case of the average number of keystrokes used or time

spent across the different modes of MTQE, there were no significant differences per type of MTQE. However, there were significant differences when the MTQE score level was higher (the higher the score level, the less time was spent, and fewer keys were typed regardless of the MTQE type). They concluded that displaying MTQE scores was not necessarily better than displaying no scores. As we shall see later in this paper, their conclusion seems to contradict our findings. This could also be due to the nature of the task, as we focused on technical translations, whereas the experiment done by Teixeira and O'Brien [16] focused on news texts. As we will see later on, our results suggest that MTQE could be beneficial for post-editing tasks. (In addition, it has to be mentioned that Teixeira and O'Brien [16] never published a full paper and most of the information included here is based on the authors' slides from MT Summit XVI and personal communication. For this reason it is impossible to know precisely all the settings of their experiment).

Finally, multiple recently published studies [18,19] confirm that PE continues to be a vibrant research topic at the intersection of translation process research, natural language processing, and human-computer interaction. Current relevant research directions include multi-modal PE [20,21], in which PE systems take several forms of input such as speech, touch, and gaze in addition to the usual keystrokes, and automatic post-editing (APE) [22–24], which attempts to automatise corrections commonly carried out in the MT output by human translators during the PE process.

To the best of our knowledge, our paper is the first attempt to measure the impact of MTQE in real scenarios accounting for the variable of getting accurate predictions and comparing against both translation from scratch and MTPE without MTQE information. In past research, experimental setups focused on accurate/inaccurate MTQE predictions. While others [7,16] investigated the impact of using accurate MTQE, the novelty of our approach lies in accounting for what happens when the machine is wrong and predicts the quality of MT output inaccurately. In addition, we use a dataset from a genre that is commonly used by professional translators.

3. Machine Translation Quality Estimation

MTQE estimates the quality of machine translation output without the need for a reference translation. This estimation can be defined differently based on the task at hand. In this research, we use the FMS as our quality metric, in line with the findings of [11,12] who identified a correlation between the editing effort and the FMS. In addition, and as also mentioned by [11,12], FMS is a score that can be understood by translators, as it is normally used in translation memories. This is in contrast to other scores like HTER and BLEU which have little meaning for translators. Following up previous research, we used the same quality threshold (75%) usually applied in the translation industry [11,12,25]. Hence, our MTQE system is trained to predict the FMS that a given MT output would have, without having access to its correct translation at any given point.

This section describes the Autodesk dataset which we used to train the MTQE approach, the MTQE approach itself, and the evaluation of the MTQE performance.

3.1. The Autodesk Data

In April 2015, Autodesk announced the release of the Autodesk Post-Editing Data corpus (http://www.islrn.org/resources/290-859-676-529-5/, accessed on 9 September 2021). It consists of parallel data where English is the source language and there are up to 12 target languages (simplified and Traditional Chinese, Czech, French, German, Hungarian, Italian, Japanese, Korean, Polish, Brazilian Portuguese, Russian, and Spanish). The size per language pair varies from 30,000 segments to 410,000 and each segment is labelled with information as to whether it comes from a TM match or it is an MT output. The post-edited target sentences are also included in the dataset. The data belongs to the technical domain, and the segments come predominantly from software manuals.

This corpus was released with MTPE tasks in mind. The fact that it is publicly available makes it a very good choice for our study, as it includes FMS which can be used to train our

MTQE. In addition, it is domain-specific making it appropriate for our user study where we try to replicate the conditions in which professional translators normally work.

For training and testing our MTQE system, we randomly selected 5500 sentences: 5000 sentences were used to train our system, and the remaining 500 sentences were kept for testing. These sentences were not used in the user study described in Section 4.

The final version of the dataset can be found in our Supplemental Materials.

3.2. Our Method

Our MTQE system is a semantically enhanced version of the QuEst++ [26] MTQE framework. We extract the 79 black box features from QuEst++ with the default settings and resources, which include a language model and vocabulary file based on the Europarl corpus. The black-features are system-independent features and are based on the source and target sentences. Therefore, they are ambivalent to the type of MT system used. We use the sentence-level features which include, among other features, the number of tokens in the source and target sentences and their ratio, the language model probability of the source and target sentences, the ratio of punctuation symbols and the ratio of percentage of numbers (A full list of features can be found at www.quest.dcs.shef.ac.uk, accessed on 9 September 2021).

Our system uses three more features in addition to those extracted from QuEst++. These features add semantic information provided by our Semantic Textual Similarity (STS) method [27,28]. To this end, we use the STS method to identify the most similar sentence to the input sentence in a collection of sentences for which the quality of their translation is known. We will refer to this sentence in the rest of the paragraph as SIM_SOURCE and its translation as SIM_TARGET. The features we add are:

- The similarity score between the input sentence and SIM_SOURCE;
- The known quality of the translation of SIM_TARGET, which in the case of this paper is the FMS retrieved from the dataset we use; and
- The FMS between the translation of the input sentence and SIM_TARGET.

We demonstrate these features further through the example in Table 1. The first STS feature we add is the similarity score between Source(1) and Source(2), which in this case is 4.2. The second feature is the FMS of MT(2) compared to PE(2), in this case this is 91. Finally, the third feature is the FMS of MT(1) compared to PE(2), which in this case is 60. More details about this quality estimation method can be found in [29].

Table 1. Example of STS comparison.

	Sentence *A*
Source(1)	Windows XP or Windows 7: On the Windows Start menu, click Run.
MT(1)	Windows XP o Windows 7: en el menú Inicio de Windows, haga clic en Ejecutar.
PE(1)	Windows XP o Windows 7: en el menú Inicio de Windows, haga clic en Ejecutar.
	Closest STS Match
Source(2)	Windows 8: From the Start screen, type run and click Run.
MT(2)	Windows 8: desde la pantalla de inicio, escriba Ejecutar y haga clic en Ejecutar.
PE(2)	Windows 8: en la pantalla de inicio, escriba ejecutar y haga clic en Ejecutar.
FMS	91
STS	4.2
FMS: MT(1) vs. PE(2)	61

The above 82 features (the 79 black box QuEst++ features plus the 3 STS features) are used to train a Support Vector Machine model to estimate the quality of a translation. We use an RBF kernel and train a regression model which estimates a continuous score between 0 and 100 for each sentence, which corresponds to the FMS. We optimise the values of C and ϵ through a grid-search which uses a 5-fold cross-validation method. We realise that the MTQE system we use in this research is outdated when considering current

MTQE approaches. The data used in this research was collected when QuEst++ was a very strong baseline for MTQE and before the advent of MTQE methods based on deep learning. We acknowledge the fact that recent MTQE methods like TransQuest [30] will perform better than our method, but we do not believe that this will dramatically influence our findings. As we stated before, our aim is to assess the impact of MTQE in a real scenario, keeping in mind that at time the predictions of MTQE will not be accurate, not assessing the quality of an MTQE method per se.

We use the Mean Absolute Error (MAE) to evaluate the performance of our system. The MAE shows how close our predictions are to the observed outcomes. For our purposes, it is a fitting way to evaluate the system's performance, as the closer we are to the observed FMS, the more useful our results would be for post-editors in real settings. While further measures of correlation would have been a useful comparison, we opted to use only MAE as the focus of our research lies more in the user study, and this evaluation was to pick the system we would use there.

3.3. Evaluation of MTQE

We trained three different MTQE systems: in the first system, we used the 79 features extracted using QuEst++ with its default language resources and without any additional semantic information provided by our STS method. For the second system, we tuned QuEst++ to our dataset by replacing some resources with Autodesk-specific data. We replaced both the English and Spanish corpora with 67,030 sentences from the Autodesk Translation Memory data. We also built a new Language Model and vocabulary file using this aligned corpus. This tuned QuEst++ to our specific domain and, as the results show, improved the performance of the baseline features.

The third system is the domain tuned version of QuEst++ (second system) enhanced with our STS features. In order to calculate the three STS features, we searched the remaining unused sentences in the Autodesk dataset for sentences with a high similarity to the 5500 randomly chosen dataset. The results are summarised in Table 2.

Table 2. MAE predicting the FMS for the Autodesk Data (low MAE indicate better performing system).

System Description	MAE
QuEst++ – out of the box	9.82
QuEst++ – tuned for in-domain data	9.78
QuEst++ – tuned for in-domain data with STS features	9.52

Examples 1–4 are taken directly from the Autodesk dataset, and show the predicted FMS (provided by our MTQE system) vs. the observed FMS (obtained by comparing the MT suggestion to the post-edited reference translation provided by Autodesk). Examples 1–3 show the prediction matching quite closely with the observed FMS. In the first two cases, MTQE would advise the translator to post-edit rather than translate from scratch, and in both cases it is the correct course of action judging by the observed FMS, as the changes are minimal. In the case of Example 4, the predicted score differs from the observed score. The difference is large enough that the MTQE system will suggest the wrong course of action to the post-editor.

Example 1. *Sample Prediction 1—Good QE-FMS: 88.661 (Predicted)/88.000 (Observed)*

a. Source: To Navigate the Marking Menu Selections.
b. MT: Para navegar por las selecciones del menú de comandos frecuentes
c. PE: Para desplazarse por las selecciones del menú de comandos frecuentes

Example 2. *Sample Prediction 2—Good QE-FMS: 87.03 (Predicted)/85 (Observed)*

a. Source: Stylizes each point based on the normal of the point
b. MT: Stylizes cada punto según la normal del punto.

c. *REF: Aplica un estilo a cada punto en función de la normal del punto.*

Example 3. *Sample Prediction 3—Good QE-FMS: 59.95 (Predicted)/54 (Observed)*

a. *Source: Create better buildings with intelligent 3D model–based design.*
b. *MT: Crear mejores edificios con modelos 3D avanzados basados en diseño.*
c. *REF: Cree mejores construcciones gracias al diseño basado en modelos 3D. 0 54 59.9523*

Example 4. *Sample Prediction 4—Bad QE - FMS: 90.88 (Predicted)/55 (Observed)*

a. *Source: For example, intensity, normal, abstractname or classification data may not be available with a point cloud.*
b. *MT: Por ejemplo, la intensidad, normal o datos de clasificación pueden no estar disponibles con una nube de puntos.*
c. *REF: Por ejemplo, es posible que los datos abstractname de intensidad, normales o clasificación no estén disponibles en una nube de puntos*

The results summarised in Table 2 show that the system performs better when tuned for the specific dataset, and its performance is improved further when augmented with STS features. However, the extent to which MTQE can be useful in a real-world setting cannot be inferred from a MAE score. This aspect is evaluated by our user study presented in the next section.

4. The User Study

The previous section presented a MTQE system trained to predict the FMS for a translated sentence. In order to understand whether this information can be useful in a real-world scenario, we replaced the predicted FMS with a binary label that would tell the translator whether they should post-edit or translate from scratch. We chose an FMS of 75 to be the threshold for post-editing, as per the findings of [12]. When we repeat the same process on the real FMS as contained in the Autodesk dataset the accuracy of the labels determined automatically is 85%. In a real-world setting, we posit that this accuracy would be of great use to translators and hope it would speed up the post-editing process. However, the extent to which this improves the efficiency of the post-editing process would need to be observed in a controlled environment with actual translators. The following subsection presents PET, the tool used in our experiment. Section 4.2 details the setting of the user study, followed by information about the professional translators involved in the study.

4.1. PET: Post-Editing Tool

We chose to use PET (https://github.com/wilkeraziz/PET, accessed on 9 September 2021) [31] as our post-editing tool. PET is a CAT tool that provides a user interface that facilitates the post-editing and translation of texts. PET is easy to use and light on extraneous features. Most importantly to us, PET records the translators' activities as they use the tool, saving information such as the time it took to complete the translation, the number of attempts, the number of keystrokes required to make the translation, and even the type of keys used. While PET is not usually the post-editors' first choice in a real-world professional post-editing situation, the fact that it collects all this useful data makes it ideal for our research. Furthermore, PET is open-source and written in Java. This allowed us to easily modify the tool to include the traffic lights described in Section 4.2. Despite the fact that translators are more familiar with and might prefer tools such as Trados Studio (https://www.trados.com/products/trados-studio/, accessed on 9 September 2021) or MemoQ (https://www.memoq.com, accessed on 9 September 2021), the lack of malleability and customisation that these tools provide did not suit our purposes for this study.

4.2. Settings of the User Study

For our user study we designed the aforementioned traffic light system using PET to provide translators information about the sentences they have to translate. The translators were presented with a sentence (in English) to translate into Spanish. Some of these sentences had automatically generated Spanish translations extracted from the Autodesk data and could be used by the translators to post-edit. All sentences are presented with one of four traffic lights, detailed below. The background colour of the PET environment changes based on the following scenarios:

- Yellow: In this case, no MT suggestion is provided and the translator has no choice but to translate from scratch.
- Blue: In this case, a MT suggestion is included but no MTQE information is presented. The translator must assess the MT suggestion and decide whether to translate or post-edit.
- Green: This indicates that the MTQE system has assessed the MT suggestion and found it merits post-editing, as it has predicted a fuzzy match score of 75 or more.
- Red: This indicates that the MTQE system has assessed the MT suggestion and found that it does not merit post-editing, as it has predicted a fuzzy match score of less than 75.

Figures 1 and 2 show how the colour coding system was displayed to the translators.

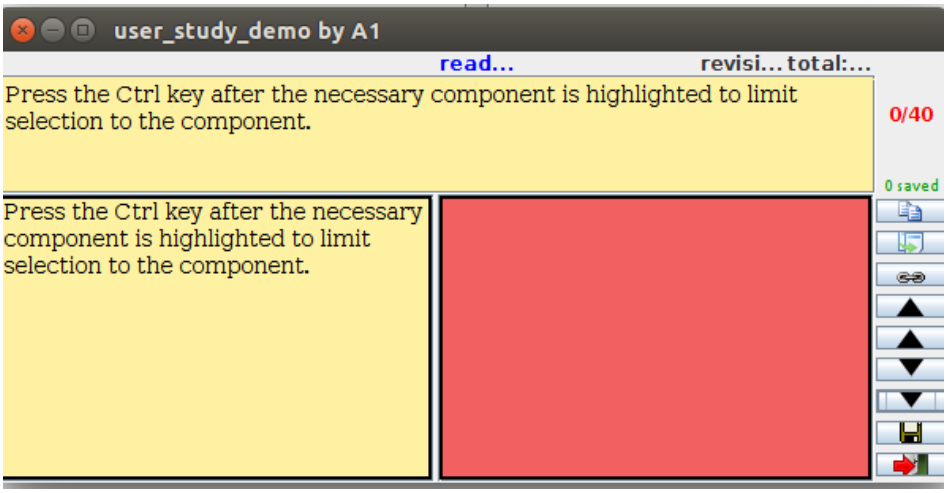


Figure 1. Translate from Scratch.

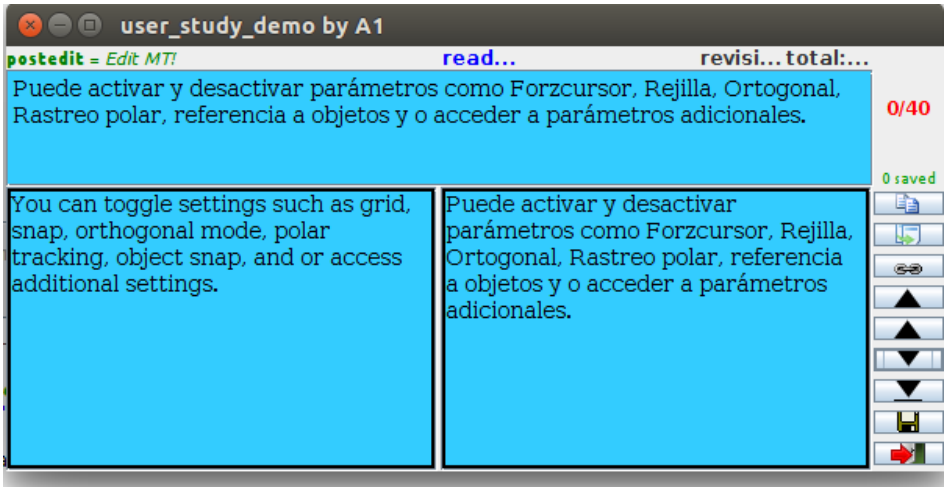


Figure 2. Postedit without MTQE.

The categories above were determined on the basis of the predicted MTQE scores. We know that some of these scores are not always precise. For this reason a different way of looking at the data is to divide it into four categories: *No MT, No QE, Good QE* and *Bad QE*. These categories are summarised in Table 3. The *Good QE* and *Bad QE* categories both provide an MTQE suggestion to the translator. However, the *Good QE* category is made up of sentences for which the MTQE system predicted a FM score close to the observed FM score (within 10% of the observed score). The *Bad QE* category consists of sentences where the MTQE system did not perform as accurately, and suggested a score that diverged from the observed FM score. The translator will not be told which sentences are *Good QE* and which are *Bad QE*, but instead be prompted to post-edit or translate from scratch based on the results of our MTQE system without any accuracy filtering. This was done to mimic a realistic scenario, where the MTQE system may not be fully accurate. Which predictions were accurate and which were wrong are only known by the researchers. The aim of this categorisation is to show us the effect of good MTQE specifically on the time and technical effort, in terms of keystrokes, as well as to try to give us insights into what happens when the translators are given inaccurate indications.

Table 3. Data Categorisation.

Label	Description	Sentences
No MT	No MT suggestion.	65
NO QE	MT Suggestion but no QE suggestion	65
Good QE	QE suggestion within 10% of observed FM score	65
Bad QE	QE suggestion more than 10% different to observed FM score	65

For the purpose of this study we selected 260 sentences (about 3000 words). As a rule of thumb, this number represents a day's work for the average professional translator (https://blog.taus.net/translation-productivity-revisited, accessed on 9 September 2021) [25]. This allows us to emulate a real-world setting by asking the translators to complete the task in one day. In addition, we wanted to investigate the impact of each type of sentence presented in Table 3. For this reason, when we extracted sentences from the Autodesk dataset, we ensured that we had 65 sentences for each category. Our final sentence count is broken down in Table 3.

4.3. The Translators

Before we ran our real study, we piloted the study with three native Spanish speakers. Even though they were not professional translators this pilot study allowed us to ensure that the real study was carried out with minimal problems. We do not report any results from this pilot study, as the participants were not professional translators, and its purpose was to test the experimental setting and confirm that the instructions provided were clear and no unforeseen issues arose. Despite taking these precautions, we still experienced a few issues, namely due to the operating system used by one of the translators (a Mac OS), and by another who was too tech-savvy and tried to use PET the way a commercial CAT tool works. Despite these issues, we ran the experiment and the results are presented in the next section.

We enlisted the help of four professional Spanish translators, each with a working proficiency of English, and a range of experience of between 3 and 14 years. All had some experience with Post-Editing tools. We asked all the translators to fill out a questionnaire detailing their professional experience. Table 4 summarises the details.

The translators had varying degrees of experience with post-editing tools, ranging between as little as one year to as much as four years. However, none of the translators had used PET before participating in this experiment. Therefore, we provided detailed instructions and a short user manual which included screenshots of the tool and directions for use. With these instructions, we hoped to familiarise the translators with the new tool before they started using it. We bundled our modified version of PET with the instructions

and sentences and emailed them to the translators. As the workload simulated a day's work for the average translator, we asked the participants to complete the task over the course of one day. All translators were compensated for their time.

Table 4. Translator Details Summarised.

	C	M	V	S
Experience in technical domains (years)	14	6	3	6
Experience as a professional translator (years)	14	6	3	9
Experience with post-editing tools (years)	2	4	3	1
Opinion of Computer-Assisted Translation tools	+	+	+	+
Opinion of post-editing tasks	−	+	+	+

5. Results and Analysis

PET records all of the operations carried out by the translators. These operations are saved in an XML file which in turn can be used to analyse the translation process. In this section we analyse the operations carried out by our four translators in an attempt to measure any increase in their productivity as a result of providing them with information about the quality of the translation.

We took the post-editing times and keystrokes from all four translators and normalised them to eliminate the effect of sentence length on the results. This was achieved by dividing each set of values by the number of tokens in the final post-edited target sentence. We chose to normalise according to target rather than source, as the sentence lengths vary greatly between Spanish and English.

Revision information, which tracks the number of times a subject returns and re-translates a sentence, was not available for subject "M" so we imputed their results with the modal value (1, which occurred 96% of the time). One observation was removed for apparently taking 30 min to complete (median translation was 28 s). Furthermore, there was a small amount of missing data on timings (see Table 5), and more cases where no keystrokes were recorded (see Table 6) (No keystrokes were recorded for 14 (3%) of sentences to be translated from scratch and 47 (about 8%) of sentences to be post-edited). To make log transformations possible without losing observations we add one to every keystroke count, a transformation that does not substantively affect the results. See Model 3 in Table 7 and Model 4 in Table 8 for comparison.

Table 5. Missing data on translation time, by subject.

	V	S	M	C
missing	0	2	0	1
present	260	258	260	259

Table 6. Cases where no keystrokes were recorded, by subject.

	V	S	M	C
non-zero	248	241	258	232
zero	12	19	2	28

Table 7. Models for post-editing keystroke count. Model 3 removes all observations where no keystrokes were recorded.

	Model 1	**Model 2**	**Model 3**
(Intercept)	−2.01 (0.73) **	−1.13 (0.80)	0.40 (0.62)
subject: S	1.84 (0.91) *	1.78 (0.91)	2.66 (0.71) ***
subject: M	1.07 (0.95)	1.08 (0.94)	−0.14 (0.71)
subject: C	2.33 (0.89) **	2.45 (0.89) **	2.38 (0.69) ***
log tokens	1.69 (0.23) ***	1.36 (0.26) ***	1.14 (0.20) ***
revisions	1.07 (0.29) ***	1.09 (0.29) ***	0.45 (0.23) *
Bad	0.21 (0.12)	−1.26 (0.74)	−1.04 (0.57)
Good	−0.49 (0.12) ***	−2.31 (0.82) **	−1.28 (0.64) *
subject: S × log tokens	−0.80 (0.32) *	−0.78 (0.32) *	−1.08 (0.25) ***
subject: M × log tokens	−0.44 (0.33)	−0.44 (0.33)	−0.10 (0.25)
subject: C × log tokens	−1.26 (0.32) ***	−1.30 (0.32) ***	−1.22 (0.25) ***
log tokens × Bad		0.53 (0.26) *	0.44 (0.20) *
log tokens × Good		0.66 (0.29) *	0.30 (0.23)
R^2	0.29	0.29	0.33
Num. obs.	588	588	541

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.

Table 8. Models for from scratch translation keystroke count. Model 3 is fitted leaving out Subject D. Model 4 removes cases where no keystrokes were recorded.

	Model 1	**Model 2**	**Model 3**	**Model 4**
(Intercept)	2.96 (0.30) ***	2.51 (0.41) ***	2.23 (0.31) ***	2.69 (0.34) ***
subject: S	−0.27 (0.11) *	0.04 (0.51)	−0.27 (0.10) **	0.45 (0.43)
subject: M	−0.27 (0.11) *	−0.04 (0.50)	−0.27 (0.10) **	−0.07 (0.42)
subject: C	−0.62 (0.11) ***	0.66 (0.52)		0.98 (0.44) *
log tokens	0.73 (0.08) ***	0.90 (0.14) ***	0.99 (0.08) ***	0.90 (0.11) ***
revisions	0.10 (0.19)	0.09 (0.19)	−0.02 (0.20)	−0.14 (0.16)
Bad	−1.80 (0.10) ***	−1.78 (0.10) ***	−1.47 (0.10) ***	−1.60 (0.08) ***
Good	−1.46 (0.11) ***	−1.43 (0.11) ***	−1.03 (0.11) ***	−1.23 (0.09) ***
subject: S × log tokens		−0.12 (0.20)		−0.25 (0.16)
subject: M × log tokens		−0.09 (0.19)		−0.07 (0.16)
subject: C × log tokens		−0.49 (0.20) *		−0.56 (0.16) ***
R^2	0.64	0.65	0.67	0.69
Num. obs.	450	450	339	436

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.

The rest of the analysis is structured as follows. It first presents an analysis of the productivity of the translators measured using the time it took them to produce the translations and the number of keystrokes required for this (Section 5.1). The effect of good and bad MTQE information on the post-editing is analysed in Section 5.2, whilst the effect of the FMS on the post-editing information is presented in Section 5.3. We compare the resulting translations in Section 5.4. The section finishes with discussion and analysis of the translators and their feedback in Section 5.5.

5.1. Analysis of Productivity

Figure 3 shows the normalised time, that each translator spent on a given type of task (translating from scratch—"*No MT*"; raw post-editing—"*No QE*"; and post-editing with MTQE information—"*QE*"). The sentences that needed to be translated from scratch (*No MT*) took the most time across all translators (on average 3.8 s per word). This means that providing automatic translation for sentences can considerably boost translator efficiency. This in itself is not an unexpected result, as MT is widely used in the translation industry nowadays to reduce the translating effort. However, for us the more interesting results are the differences between post-editing with MTQE and without MTQE. We take a closer look at the bars marked "*No QE*" and those marked "*QE*". The normalised (per word) number of seconds drops from an average of 2.9 s to 2.4. This indicates that MTQE cuts post-editing time by an average of 0.5 s per word.

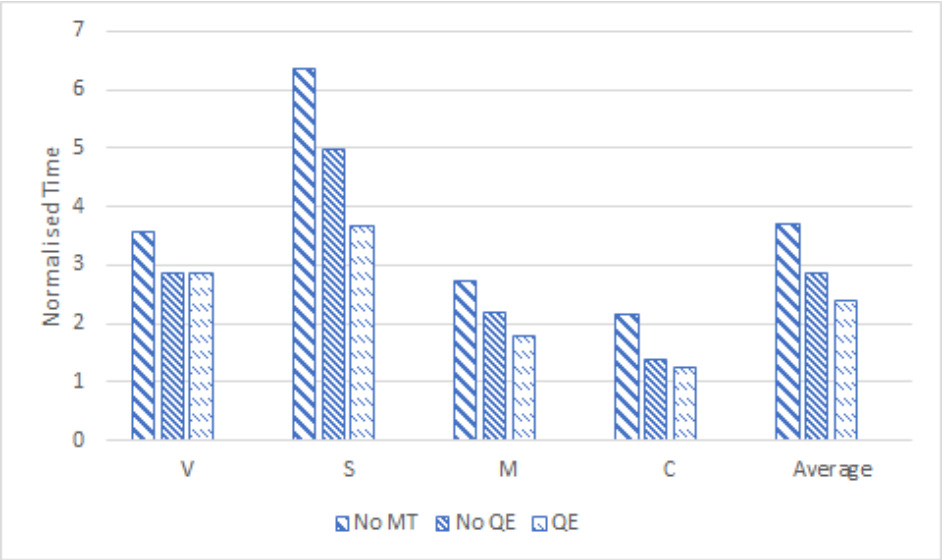


Figure 3. Normalised Number of seconds per word spent translating/post-editing.

Figure 4 analyses the activity of translators from the point of view of the number of keystrokes per word, based on the type of task (translating from scratch—"*No MT*"; raw post-editing—"*No QE*"; and post-editing with QE information—"*QE*"). This helps us measure the effort in addition to the time saved. As expected, the number of keystrokes used in the "*No QE*" and "*QE*" conditions in Figure 4 is clearly lower than the number of keystrokes used when translating from scratch (about 4 s per word). The same observation carries over when measuring the difference in keystrokes for post-editing with and without MTQE. The average number of keystrokes drops from 3.67 for "*No QE*" to 2.25 for "*QE*". This suggests that the inclusion of MTQE cuts post-editing effort by 0.4 keystrokes per word. The reduction of the number of keystrokes between the setting where no machine translation is provided and those where a translation was available is much greater than the reduction of the time between the same settings. This is to be expected given that even when a good automatic translation is available, translators need to spend time to read it and assess its quality.

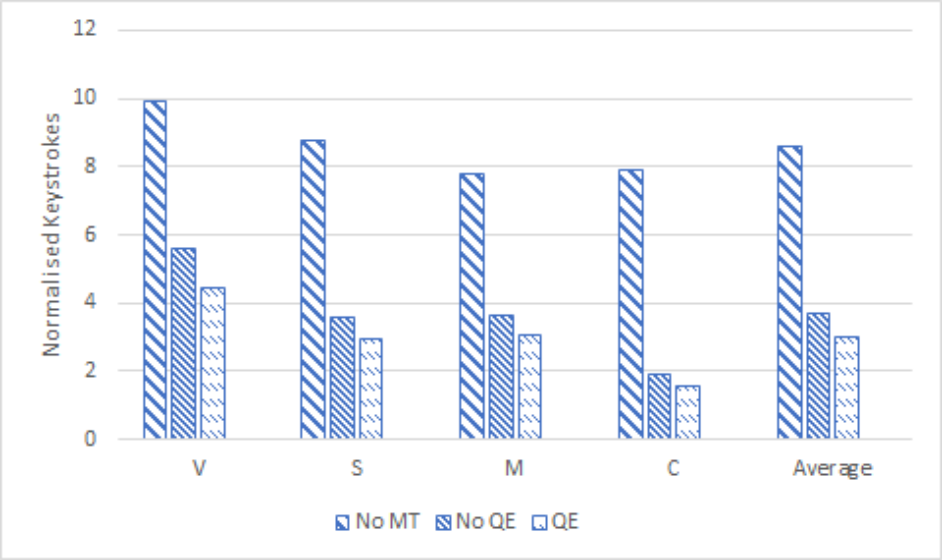


Figure 4. Normalised keystrokes spent translating/post-editing.

5.2. The Effect of Good QE vs. Bad QE on Post-Editing Efficiency

We compare translation times under two conditions, "*Good QE*" and "*Bad QE*" to a "*No QE*" baseline. Figure 5 plots the raw data (Full replication code for analysis can be found at https://github.com/conjugateprior/hbtranslation, accessed on 9 September 2021). Since we can operationalise effort either in translation time (in seconds) or in 'effort' (in keystroke count) we present models of both dependent variables. We also expect effects to to be multiplicative, so work with logged time in seconds and logged keystroke count. As noted above, in a few cases there are zeros in these dependent variables so we add 1 before logging, but results are not sensitive to leaving out zero outcomes. In an attempt to improve estimation precision we also control for the length (in tokens) of the correct translation. We use unadjusted log length in tokens since this is always positive.

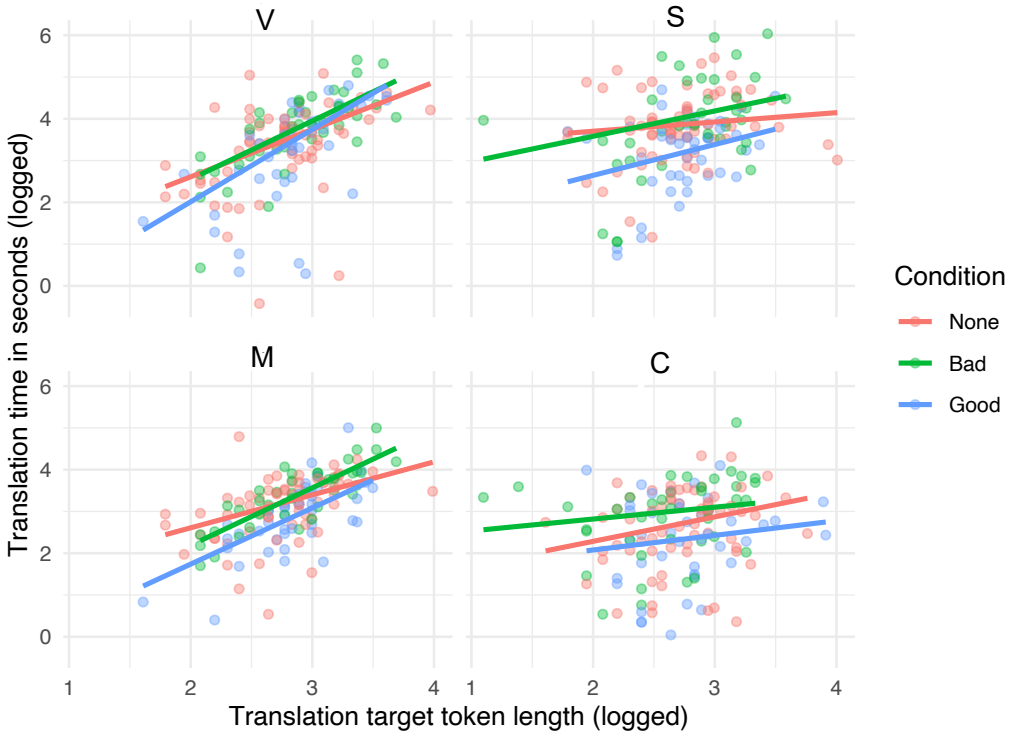


Figure 5. Summary of Post editing times by subject and condition (Good vs. Bad)—Raw Data.

The most parsimonious models for this part of the data set were linear models containing categorical variables representing condition and subject, log target sentence length in tokens, the number of times the subject revisited the translation task, plus interactions between subject and log target tokens, and in the second model between log target tokens and condition. These interactions are to be expected when the size of the translation task differentially affects individual translator times (due to individual differences) or the effects of different quality of advice (heterogenous treatment effects).

Since evidence for differential treatment effects are relatively weak (though statistically significant) we present these models in Tables 7 and 9. In these tables, classical standard errors are noted in parentheses. However, since our primary interest is in the the effect of each condition, Figure 6 plots the marginal effect of condition in terms of pairwise differences. (Since marginal effects average over the interactions our design means that they are almost identical across models, so we show effects from Model 1.) These marginal effects are corrected for multiple comparisons using Tukey's method.

From the marginal effects we see that "*Good QE*" speeds translation time significantly relative to "*No QE*", and also relative to "*Bad QE*"."*Bad QE*" does not significantly improve translation times relative to"*No QE*", although we are less confident of this because the significance fluctuates either side of the 0.05 level depending on model specification.

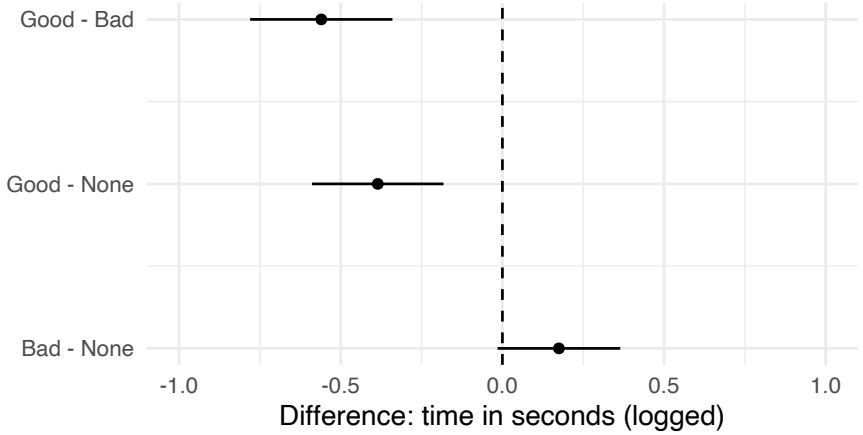


Figure 6. Differences between condition, averaging over covariates and interactions. Confidence intervals are simultaneous and corrected for multiple comparison using Tukey's method.

Table 9. Models for post-editing time (in log seconds). Model 2 allows advice quality to have different effects on translation time depending on target sentence length.

	Model 1	**Model 2**
(Intercept)	−1.18 (0.52) *	−0.43 (0.57)
subject: S	2.22 (0.65) ***	2.19 (0.65) ***
subject: M	0.40 (0.67)	0.40 (0.67)
subject: C	2.03 (0.63) **	2.15 (0.63) ***
log tokens	1.41 (0.16) ***	1.13 (0.19) ***
revisions	0.64 (0.22) **	0.65 (0.22) **
Bad	0.17 (0.09) *	−1.15 (0.53) *
Good	−0.39 (0.08) ***	−1.87 (0.58) **
subject: S × log tokens	−0.68 (0.23) **	−0.67 (0.23) **
subject: M × log tokens	−0.23 (0.24)	−0.24 (0.23)
subject: C × log tokens	−1.00 (0.23) ***	−1.04 (0.23) ***
log tokens × Bad		0.48 (0.19) *
log tokens × Good		0.53 (0.21) *
R^2	0.35	0.36
Num. obs.	586	586

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.

5.3. The Effect of the FMS on Post-Editing Effort

To gain further insight into the results in Section 5.2, we take a closer look at the sentences in the "Good QE" and "Bad QE" categories, and their range of FMS.

Therefore, we divide the sentences in our data into four categories:

- *GoodQE Translate*: (21 sentences) The observed FM score is <75, and the user is given a red light.
- *GoodQE Postedit*: (42 sentences) The observed FM score is >75, and the user is given a green light.
- *BadQE Translate*: (25 sentences) The observed FM score is >75, and the user is given a red light.
- *BadQE Postedit*: (40 sentences) The observed FM score is <75, and the user is given a green light.

We therefore attempt to control for the differences in FMS. Table 10 shows a breakdown of the number of sentences by FMS range and the labels shown to translators (*Postedit* or *Translate*).

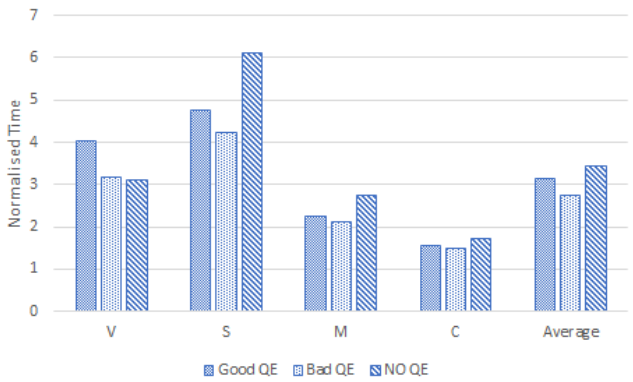
Table 10. Number of Sentences by Range.

FM Range	≤75	(75–100]
Postedit	25	42
Translate	21	40
Total	46	82

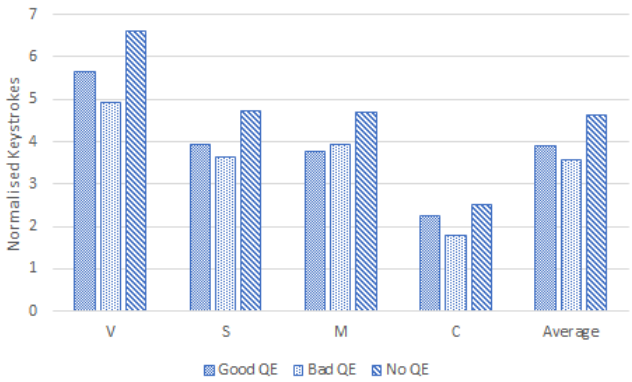
In Figures 7 and 8 we organise the sentences by their observed Fuzzy Match scores. The "Good QE" and "Bad QE" categories indicate whether the translator was given the correct or incorrect prompt. For comparison, we also include the "*No QE*" category as a control group. We also include the average normalised time and keystrokes over all four translators for each category. Figure 7a,b shows the normalised time and keystrokes (respectively) for each translator spent on sentences with FMS $\leq$ 75. Here a correct label ("*Good QE*") would indicate translating from scratch, and an incorrect label ("*Bad QE*") would indicate attempting to post-edit. On average, it seems that in the case of these particular sentences, attempting to post-edit still reduced the time and overall effort overall, despite the low accuracy of the quality estimation. This is especially pronounced for Translator "V" and Translator "S". This might indicate that these translators were more likely to follow the traffic light' suggestion.

Things are much less defined in the case of translations with higher FMS. Here, the MTQE suggestions have a less significant impact on translator time and effort. Figure 8a shows the normalised time spent post-editing sentences with a FMS $>$ 75. With the exception of Translator "S", none of the translators show a significant difference in time between "Good QE" and "Bad QE". The same can be observed for effort in Figure 8b.

Our results show that "*Good QE*" improves the efficiency of translators over "*No QE*" in most cases. However, in cases with an FMS $\leq$ 75, the quality of the QE (Whether it is "*Good QE*" or "*Bad QE*"), matters much less. This seems to suggest that MTQE is most helpful in cases of low MT quality, where the decision to post-edit or translate from scratch is difficult. This is also in line with [7], who found that the improvements in efficiency are only statistically significant for instances where HTER $>$ 0.1.

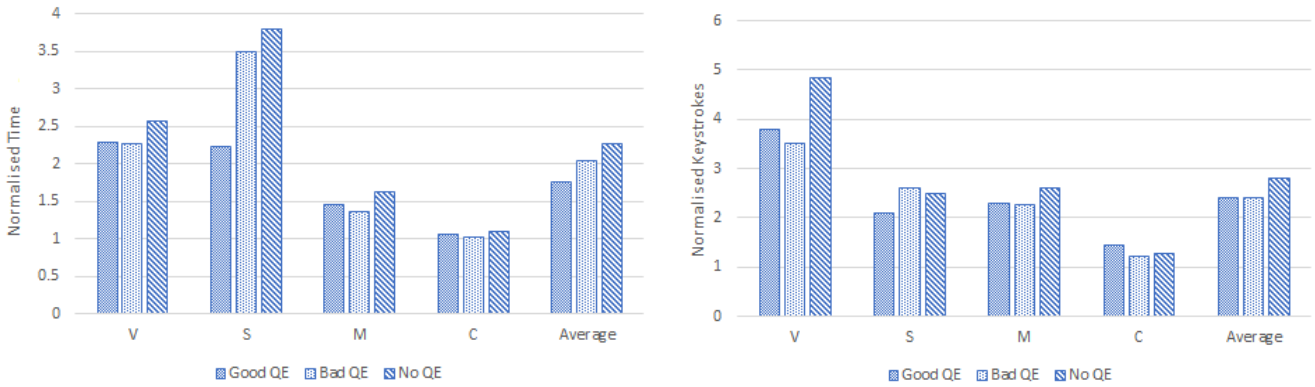


(**a**) Normalised number of seconds per word spent translating/post-editing



(**b**) Normalised number of keystrokes per word spent translating/post-editing

Figure 7. Sentences with FMS $\leq$ 75.



(**a**) Normalised number of seconds per word spent translating/post-editing

(**b**) Normalised number of Keystrokes per word spent translating/post-editing

Figure 8. Sentences with FMS $>$ 75.

5.4. Quality of the Translation

In order to find out whether there are differences between the resulting translations, we take a look at the FMS scores of the translators' sentences, comparing them to the post-edited reference provided by Autodesk. In the context of this paper, we report the results using FMS for comparison in Figure 9. In addition to FMS, we also experimented with BLEU, METEOR, HTER, and word counts for comparison. The results obtained using those metrics are very similar to those reported here and for this reason are not included in the paper.

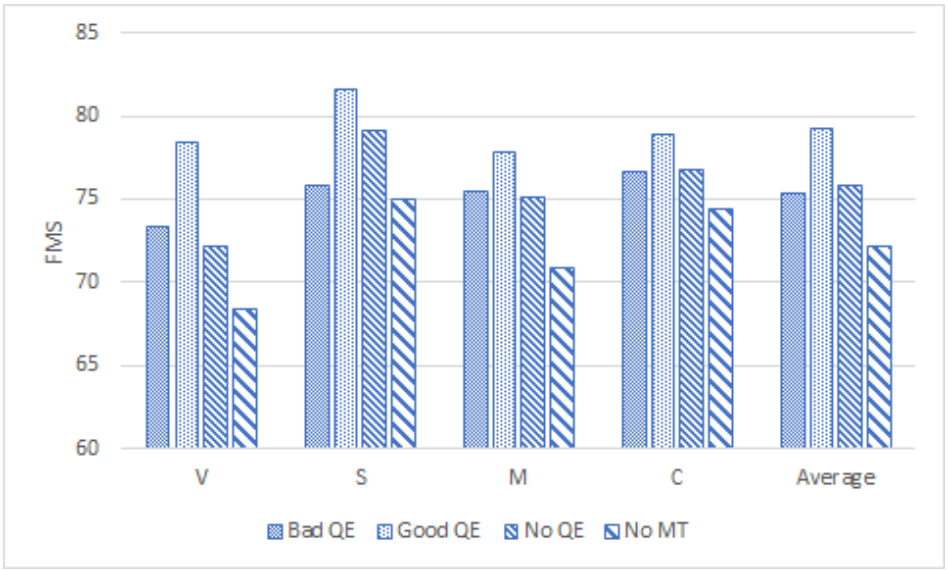


Figure 9. FMS scores for post-edited sentences compared to gold reference translation.

We find that despite their varying levels of experience, all four translators achieved fairly high scores. We found that the FMS vary the most for the "*No MT*" category, with an average standard deviation of 7.6. This result was anticipated, as we expect the results to vary the most when the translators do not have an MT suggestion as a starting point. Furthermore, they are not Autodesk usual translators and are not familiarised with the terminology and style required by Autodesk. As their MT engine is deployed in-house, it is expected that it mimics the style of their translators and uses the right terminology. Examples 5 and 6 show two such examples.

Example 5. *Source: From point, Next point*
Gold: Desde el punto, Siguiente punto
MT: Punto inicial, punto siguiente
Post-edited translations:

a. Punto inicial, punto siguiente.
b. Punto inicial, punto siguiente.
c. Desde el punto, Siguiente punto.
d. Punto de origen, Punto siguiente.

Example 6. *Source: This is required for a Core-shift analysis.*
Gold: Esto es necesario para un análisis de desplazamiento del macho.
MT: Esto es necesario para una Core-shift análisis.
Post-edited translations:

a. Esto es necesario para un análisis Core-shift.
b. Esto es necesario para un análisis de cambio de núcleo.
c. Esto es necesario para un análisis de desplazamiento de núcleo.
d. Se requiere para un anàlisis de Core-shift.

We also observe that the category with the highest FMS scores is the GoodQE category. This is consistent across all translators and on average. As this is the category with the highest improvement in post-editing time and effort, we can safely conclude that this change is not linked to lower levels of target quality.

5.5. Analysis of the Translators

In Table 4, we summarised the translators' years of experience both as professional translators and post-editors. When we compare these findings to the results in Figure 3, we find a negative correlation between the years of experience and the time spent both translating and post-editing. The exception is Translator "S", who despite 9 years as a professional translator, spends over twice as much time per word as the rest of the translators. This could reflect the lack of experience that Translator "S" possesses in terms of post-editing tools, or at least with PET. Despite the discrepancy in time, the keystrokes of Translator "S" do not seem to vary that much from the rest of the translators in terms of effort (keystrokes in Figure 4). This lends more weight to the theory that the time spent getting familiar with the PET tool is responsible for the additional time observed in the results.

To gain a better understanding of our translators and post-editing tools, all translators were asked to fill out questionnaires before and after the task. We summarise their answers in Table 11, While all four translators approved of the MT suggestions, they all experienced difficulties with navigating the Post-editing Tool, which may have affected both their results and opinions of Quality Estimation. Despite the results above, only one of the four translators felt that Quality Estimation was helpful, mentioning that they liked getting a first impression via the traffic light system. This translator, Translator "V", had the fewest years of experience both with post-editing tools and translating in general. MT suggestions were helpful for three out of four translators, but one insisted that translating from scratch was better, despite the conflicting results included above where a high increase in efficiency was observed. In fact, the results above show that all translators benefited from MTQE, despite their impressions. In particular, Translator "S", whose normalised time was cut from 4.9 to 3.7 when MTQE information was included. While Translator "V" did not gain much in terms of time, MTQE-informed post-editing cut down the effort (in terms of keystrokes) by 1 keystroke per word.

Table 11. Professional Translators and their Opinions after participating in the user study.

Translator Opinion	V	S	M	C
Professionalism of Task	Yes	No	No	Yes
MT Quality	Good	Good	Good	Good
Usefulness of MTQE	Good	Bad	Bad	Bad
Accuracy of MTQE	Good	Bad	Bad	Bad
Opinion of MTQE	Positive	Negative	Negative	Negative

6. Conclusions

We designed and implemented a user study to investigate the impact of using MTQE information in the post-editing workflow. To achieve this, we ran a study using 260 sentences from the Autodesk Post-editing Parallel Corpus. We used our own version of Quest++ enhanced with semantic features to estimate the quality of the MT provided for each sentence. The estimation is calculated as fuzzy match score. This information is used to inform translators whether they should translate a sentence from scratch or post-edit it. The semantically enhanced QuEst++ system, trained, and tuned on Autodesk data, performed better than other other variants we tried on our test set. The observed MAE for the Autodesk data corpus is of 9.5 and we obtain an accuracy of 85% for the labels *Translate* and *Postedit*. Furthermore, we conducted a user study to investigate the impact of using the MTQE information in the post-editing workflow. After a pilot study with 40 sentences to test our setup, we invited four professional translators to take part in a post-editing/translation task, using a traffic light system to provide MTQE information. The translators were asked to use a tool which records their time and all their editing operations. All four translators were paid for their work.

Our results show that MTQE can have a positive impact on the efficiency of the translator workflow when the MT suggestion is low quality (FMS $\leq$ 75). Furthermore, we also show that good MTQE can cut translating time and effort significantly regardless of the quality of the MT suggestion. This seems to contradict the findings of Teixeira and O'Brien [16], who find no significant difference in time or effort for post-editing for MTQE. As we mentioned before, these different results can be explained by the fact that they used newspaper articles. Translator feedback still seems quite negative in spite of this improvement which suggests a better post-editing tool might be required.

Our study still faces several limitations. The small number of sentences (260 overall) makes the data highly sensitive to outliers. A larger dataset, post-edited over several days, might give us a better picture of whether or not the results are significant. Furthermore, due to limited resources, we were only able to use four translators. A larger pool of translators would have allowed us to further investigate the impact experience on the effectiveness of MTQE in the translation workflow. The use of Quest++ as MTQE tool is another limitation of the study. It would be interesting to repeat the experiments using a state of the art MTQE method to find out to what extent the findings are influenced by the accuracy of the MTQE tool.

Our conclusions identify several avenues of future work. The current study includes four professional translators and 260 segments to translate. This reflects a full day's work for a professional translator. While expanding the scope of translators and segments would be quite challenging and expensive, it would provide a more robust and conclusive picture of the effect of MTQE. Furthermore, one important avenue of future work would be to test whether the results can be replicated for other language pairs and domains. Our experiments are limited to the English-Spanish language pair in the technical domain. Similar findings in experiments with other language pairs would demonstrate the need for accurate and reliable MTQE, as well as the need to integrate it in professional translation workflows to improve post-editing efficiency.

Supplementary Materials: The dataset produced in this research is available at https://dinel.org.uk/projects/postediting-dataset/.

Author Contributions: Conceptualization, C.P.E.; methodology, H.B.; C.O. and C.P.E.; software, H.B. and C.O.; formal analysis, H.B.; C.O.; C.P.E.; M.Z. and W.L.; resources, C.O.; data curation, H.B.; writing—original draft preparation, H.B.; C.P.E.; writing—review and editing, C.O.; C.P.E. and M.Z.; supervision, C.O.; M.Z.; funding acquisition, C.O. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially supported by the People Programme (Marie Curie Actions) of the European Union's Framework Programme (FP7/2007–2013) under REA grant agreement n° 317471 and Science Foundation Ireland in the ADAPT Centre (Grant 13/RC/2106) (www.adaptcentre.ie, accessed on 9 September 2021) at Dublin City University.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

FMS	Fuzzy Match Score
MTQE	Machine Translation Quality Estimation
STS	Semantic Textual Similarity

References

1. Zaretskaya, A.; Pastor, G.C.; Seghiri, M. Translators' Requirements for Translation Technologies: A user survey. In Proceedings of the 7th International Conference of the Iberian Association of Translation and Interpreting Studies (AIETI), New Horizons in Translation and Interpreting Studies, Malaga, Spain, 29–31 January 2015; pp. 133–134.
2. Schneider, D.; Zampieri, M.; van Genabith, J. Translation Memories and the Translator: A report on a user survey. *Babel* **2018**, *64*, 734–762. [CrossRef]
3. Koponen, M. Is machine translation post-editing worth the effort? A survey of research into post-editing and effort. *J. Spec. Transl.* **2016**, *25*, 131–148.
4. Parra Escartín, C.; Béchara, H.; Orăsan, C. Questing for Quality Estimation A User Study. *Prague Bull. Math. Linguist.* **2017**, *108*, 343–354. [CrossRef]
5. Guerberof, A. Productivity and Quality in the Post-Edition of Outputs from Translation Memories and Machine Translation. Ph.D. Thesis, Rovira and Virgili University, Tarragona, Spain, 2014.
6. Zampieri, M.; Vela, M. Quantifying the Influence of MT Output in the Translators' Performance: A case study in technical translation. In Proceedings of the EACL 2014 Workshop on Humans and Computer-Assisted Translation, Gothenburg, Sweden, 26 April 2014; pp. 93–98.
7. Turchi, M.; Negri, M.; Federico, M. MT Quality Estimation for Computer-assisted Translation: Does it Really Help? In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers), Beijing, China, 26–31 July 2015; Association for Computational Linguistics: Stroudsburg, PA, USA, 2015; pp. 530–535.
8. Federico, M.; Bertoldi, N.; Cettolo, M.; Negri, M.; Turchi, M.; Trombetti, M.; Cattelan, A.; Farina, A.; Lupinetti, D.; Martines, A.; et al. The matecat tool. In Proceedings of the COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations, Dublin, Ireland, 23–29 August 2014; pp. 129–132.
9. Snover, M.; Dorr, B.; Schwartz, R.; Micciulla, L.; Makhoul, J. A Study of Translation Edit Rate with Targeted Human Annotation. In Proceedings of the Association for Machine Translation in the Americas (AMTA), Cambridge, MA, USA, 8–12 August 2006; pp. 223–231.
10. Koehn, P.; Hoang, H.; Birch, A.; Callison-Burch, C.; Federico, M.; Bertoldi, N.; Cowan, B.; Shen, W.; Moran, C.; Zens, R.; et al. Moses: Open Source Toolkit for Statistical Machine Translation. In Proceedings of the Association for Computational Linguistics (ACL), Prague, Czech Republic, 23–30 June 2007; pp. 177–180.
11. Parra Escartín, C.; Arcedillo, M. Machine translation evaluation made fuzzier: A study on post-editing productivity and evaluation metrics in commercial settings. In Proceedings of the MT Summit XV, Miami, FL, USA, 30 October–3 November 2015.
12. Parra Escartín, C.; Arcedillo, M. Living on the edge: Productivity gain thresholds in machine translation evaluation metrics. In Proceedings of the Fourth Workshop on Post-Editing Technology and Practice, Miami, FL, USA, 30 October–3 November 2015; pp. 46–56.
13. Moorkens, J.; O'Brien, S.; da Silva, I.A.L.; de Lima Fonseca, N.B.; Alves, F. Correlations of perceived post-editing effort with measurements of actual effort. *Mach. Transl.* **2015**, *29*, 267–284. [CrossRef]
14. Moorkens, J.; Way, A. Comparing Translator Acceptability of TM and SMT outputs. *Balt. J. Mod. Comput.* **2016**, *4*, 141–151.

15. Moorkens, J.; O'Brien, S. Assessing User Interface Needs of Post-Editors of Machine Translation. In *Human Issues in Translation Technology: The IATIS Yearbook*; Kenny, D., Ed.; Routledge: Oxford, UK, 2017; pp. 109–130.
16. Teixeira, C.; O'Brien, S. The Impact of MT Quality Estimation on Post-Editing Effort. In Proceedings of the MT Summit XVI, Volume 2: Users and Translators Track, Nagoya, Japan, 18–22 September 2017; pp. 211–233.
17. Graham, Y.; Mathur, N.; Baldwin, T. Accurate evaluation of segment-level machine translation metrics. In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics Human Language Technologies, Denver, CO, USA, 31 May–5 June 2015; pp. 1183–1191.
18. Tezcan, A.; Hoste, V.; Macken, L. Estimating Post-editing Time Using a Gold-standard Set of Machine Translation Errors. *Comput. Speech Lang.* **2019**, *55*, 120–144. [CrossRef]
19. Wang, K.; Wang, J.; Ge, N.; Shi, Y.; Zhao, Y.; Fan, K. Computer Assisted Translation with Neural Quality Estimation and Auotmatic Post-Editing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, Online, 16–20 November 2020; pp. 2175–2186.
20. Herbig, N.; Düwel, T.; Pal, S.; Meladaki, K.; Monshizadeh, M.; Krüger, A.; van Genabith, J. MMPE: A multi-modal interface for post-editing machine translation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 1691–1702.
21. Herbig, N.; Pal, S.; Krüger, A.; van Genabith, J. Multi-modal estimation of cognitive load in post-editing of machine translation. In *Translation, Interpreting, Cognition*; Language Science Press: Berlin, Germany, 2021; pp. 1–32.
22. Chollampatt, S.; Susanto, R.H.; Tan, L.; Szymanska, E. Can Automatic Post-Editing Improve NMT? In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 16–20 November 2020; pp. 2736–2746.
23. do Carmo, F.; Shterionov, D.; Moorkens, J.; Wagner, J.; Hossari, M.; Paquin, E.; Schmidtke, D.; Groves, D.; Way, A. A review of the state-of-the-art in automatic post-editing. *Mach. Transl.* **2021**, *35*, 101–143. [CrossRef]
24. Lee, W.; Jung, B.; Shin, J.; Lee, J.H. Adaptation of Back-translation to Automatic Post-Editing for Synthetic Data Generation. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Online, 19–23 April 2021; pp. 3685–3691.
25. Plitt, M.; Masselot, F. A productivity test of statistical machine translation postediting in a typical localisation context. *Prague Bull. Math. Linguist.* **2010**, *93*, 7–16. [CrossRef]
26. Specia, L.; Paetzold, G.; Scarton, C. Multi-level Translation Quality Prediction with QuEst++. In Proceedings of the ACL-IJCNLP 2015 System Demonstrations; Association for Computational Linguistics and The Asian Federation of Natural Language Processing, Beijing, China, 26–31 July 2015; pp. 115–120.
27. Gupta, R.; Bechara, H.; El Maarouf, I.; Orasan, C. UoW: NLP techniques developed at the University of Wolverhampton for Semantic Similarity and Textual Entailment. In Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval'14), Dublin, Ireland, 23–24 August 2014; pp. 785–789.
28. Béchara, H.; Costa, H.; Taslimipoor, S.; Gupta, R.; Orasan, C.; Corpas Pastor, G.; Mitkov, R. MiniExperts: An SVM approach for Measuring Semantic Textual Similarity. In Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval'15, Denver, CO, USA, 4–5 June 2015; pp. 96–101.
29. Béchara, H.; Parra Escartín, C.; Orasan, C.; Specia, L. Semantic textual similarity in quality estimation. In Proceedings of the 19th Annual Conference of the European Association for Machine Translation, Riga, Latvia, 30 May–1 June 2016; pp. 256–268.
30. Ranasinghe, T.; Orăsan, C.; Mitkov, R. TransQuest: Translation quality estimation with cross-lingual transformers. In Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain, 8–13 December 2020; pp. 5070–5081. [CrossRef]
31. Aziz, W.; Castilho, S.; Specia, L. PET: A Tool for Post-editing and Assessing Machine Translation. In Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, 23–25 May 2012; pp. 3982–3987.