Neural Machine Translation Quality and Post-Editing Performance

Vilém Zouhar,[⋆] **Aleš Tamchyna**,[†] **Martin Popel**,[⋆] **Ondřej Bojar**[⋆]

Charles University, Faculty of Mathematics and Physics[⋆]
Institute of Formal and Applied Linguistics
Malostranské Náměstí 25, 118 00 Prague, Czech Republic
{zouhar, popel, bojar}@ufal.mff.cuni.cz

Memsource[†]
Spálená 108/51, Prague, Czech Republic
ales.tamchyna@memsource.com

Abstract

We test the natural expectation that using MT in professional translation saves human processing time. The last such study was carried out by Sanchez-Torron and Koehn (2016) with phrase-based MT, artificially reducing the translation quality. In contrast, we focus on neural MT (NMT) of high quality, which has become the state-of-the-art approach since then and also got adopted by most translation companies.

Through an experimental study involving over 30 professional translators for English→Czech translation, we examine the relationship between NMT performance and post-editing time and quality. Across all models, we found that better MT systems indeed lead to fewer changes in the sentences in this industry setting. The relation between system quality and post-editing time is however not straightforward and, contrary to the results on phrase-based MT, BLEU is definitely not a stable predictor of the time or final output quality.

1 Introduction

Machine translation is increasingly utilized in the translation and localization industry. One of the most common use cases is MT post-editing where human translators use MT outputs as a starting point and make edits to obtain the final translation. This process has been shown to be more efficient than translating from scratch, i.e. it and can lead to reductions in cost and delivery time (Plitt and Masselot, 2010).

While various studies have looked at how MT post-editing affects translation quality and speed in general, few have attempted to measure how the *quality* of MT outputs affects the productivity of translators. In this work, we attempt to answer the following questions:

- How strong is the relationship between MT quality and post-editing speed?
- Does MT quality have a measurable impact on the quality of the post-edited translation?
- Is the effect of MT quality still persistent in a second round of post-editing ("revision")?
- Is the post-editing process different when human translation is used instead of MT as the input?
- How large are the edits in the different rounds of post-editing?

We have carried out a large-scale study on one language pair that involved over 30 professional translators and translation reviewers, who worked in two stages to post-edit outputs of 13 different sources (11 MT engines, 1 raw source, 1 human reference). This allowed us to collect not only the post-editing times but also to estimate the quality of the produced results. Based on this data, we present an in-depth analysis and provide observations and recommendations for utilizing MT in localization workflows. While the task for humans for both rounds is the same (improve a given translation by freely editing it), we strictly distinguish the first and second rounds, using the term "post-editor" in the first round and "reviewer" in the second round.

We make the code along with all collected data (including all translations) publicly available.[1]

[1] github.com/ufal/nmt-pe-effects-2021

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 10204–10214
November 7–11, 2021.

2 Related Work

One of the earliest experiments that noticed a significant correlation between various automatic evaluation metrics and post-editing speed was performed by Tatsumi (2009). The survey of Koponen (2016) briefly covers the history of post-editing and pinpoints the two main topics: effort and final output quality. The authors conclude that post-editing improves both compared to translating the original text from scratch, given suitable conditions (good MT quality and translator experience with post-editing). Experiments were done by Mitchell et al. (2013) and Koponen and Salmi (2015) show that purely monolingual post-editing leads to results of worse quality than when having access to the original text as well. Finally, Koponen (2013) comments on the high variance of post-editors, which is a common problem in post-editing research (Koponen, 2016).

Interactive MT is an alternative use case of computer-assisted translation and it is possible that effort or behavioural patterns in interactive MT could be used as a different proxy extrinsic measure for MT quality. Post-editor productivity has also been measured in contrast to interactive translation prediction by Sanchis-Trilles et al. (2014).

Similar Experiments. Our work is most similar to Sanchez-Torron and Koehn (2016) and Koehn and Germann (2014), which served as a methodological basis for our study.

While most of the previous works focused on Statistical MT, we experiment solely with Neural MT (NMT) models. Many studies have shown that NMT has very different properties than older MT models when it comes to post-editing (Koponen et al., 2019). For instance, NMT outputs tend to be very fluent which can make post-editing more cognitively demanding and error-prone as suggested by Castilho et al. (2017). Popel et al. (2020) showed that in a specific setting, the adequacy of NMT is higher than that of human translators. We believe that the relationship between NMT system quality and PE effort is not a simple one and that older results based on statistical MT may not directly carry over to NMT. The first of the six challenges listed by Koehn and Knowles (2017) suggests that fluency over adequacy can be a critical issue: *NMT systems have lower quality out of domain, to the point that they completely sacrifice adequacy for the sake of fluency.*

Additionally, our focus is state-of-the-art NMT systems, which was not true for Sanchez-Torron and Koehn (2016), who constructed 9 artificially severely degraded statistical phrase-based MT systems. The experiment by Koehn and Germann (2014) used only 4 MT systems. Our focus is motivated by the industry's direct application: *Considering the cost of skilled staff and model training, what are the practical benefits of improving MT performance?*

In contrast to the previous setups, we evaluate two additional settings: post-editing human reference and translating from scratch, corresponding to a theoretical[2] BLEU of 100 and 0, respectively. We also consider the quality of the PE output and not only the process itself.

Sanchez-Torron and Koehn (2016) found a linear relationship between BLEU and PE effort: *for each 1-point increase in BLEU, there is a PE time decrease of 0.16 seconds per word, about 3-4%.* The performance of the MT systems they use is, however, close to uniformly distributed between 24.85 and 30.37. The observed linear relationship can then be partially attributed to the lower MT performance of artificially uniformly distributed MT systems.

Neural MT. An experiment by Koponen et al. (2020) considers 4 neural MT systems in a similar setting. The quality of these systems is below the state of the art.[3] The focus of this work was also to measure the difference between translating from scratch and post-editing, which was confirmed to be in favour of the latter. The contrast of using translation memories and NMT on little-explored language pairs was examined by Läubli et al. (2019).

3 Experiment Design

In this section, we thoroughly describe the design of the study, including the used data, MT engines and the translation process.

[2] In fact, humans never produce the same translation, so BLEU of 100 is unattainable, and the source text often contains some tokens appearing also in the output, so not translating can reach BLEU scores of e.g. 3 or 4.

[3] Document-level BLEU of 19.3 on miscellaneous FI→SV OPUS (Tiedemann, 2012) data. Current state of the art is 29.5 on the FIKSMÖ benchmark (Tiedemann et al., 2020).

3.1 Documents

In total, we used 99 source lines (segments) of 8 different parallel English documents for which Czech human reference translations were available. One line can contain more than one sentence, which is reflected by the rather high average sentence length of 25 words. We chose the domains to mirror common use-cases in localization: 36 lines of news texts (WMT19 News testset), 29 lines from a lease agreement (legal text), 23 lines from an audit document (Zouhar et al., 2020), and 11 lines of technical documentation (Agirre et al., 2015). The translators received all documents joined together in a single file, with clearly marked document boundaries. For clarity, we will refer to the whole set simply as *"file"* and the individual parts as *"documents"*.

3.2 Machine Translation Models

In total, we used 13 MT models of various quality. Models M01–M11 are based on the setup, training procedure and data of Popel (2020). We chose this particular approach because it has been reported to reach human translation quality (Popel et al., 2020). For our purposes, we reproduce the training, stopping it at various stages of the training process. All MT systems translate sentences in isolation, with the exception of M11, which is a document-level system (replicating CUNI-DocTransformer in Popel (2020)). Systems MT01–MT10 differ only in the number of training steps, which affects also the ratio of authentic- and synthetic- data checkpoints in the hourly checkpoint averaging (Popel et al., 2020): the best dev-set BLEU was achieved with 6 authentic-data and 2 synthetic-data checkpoints, but we include also models with other ratios (cf. column *ACh* in Table 1).

In addition to the internal MT system variants, we also included outputs of commercially available models by Google[4] and Microsoft.[5]

Overview of all the 13 MT systems is provided in Table 1. Although the range of BLEU scores is very large (25.35–37.44), the scores are not spread out evenly (average 34.65).[6] Most of the systems are concentrated in the upper half of the range. This better reflects realistic scenarios in localization workflows where users can typically decide among several engines of comparable but not identical performance.

Model	TER	BLEU	Steps [k]	ACh
M01	0.729	25.35	25.4	8
M02	0.678	31.61	29.0	8
M03	0.655	33.09	29.3	8
M04	0.648	33.63	33.0	8
M05	0.622	35.22	72.8	6
M06	0.624	35.68	997.1	0
M07	0.604	36.58	1015.2	5
M08	0.600	36.41	1022.4	6
M09	0.603	37.40	1055.0	8
M10	0.600	37.44	1058.6	6
M11	0.601	37.37	698.5	5
Google	0.623	37.56	–	–
Microsoft	0.632	33.06	–	–

Table 1: Overview of MT systems used. TER and BLEU were measured by SacreBLEU[7] (Post, 2018). Steps mark the number of training steps in thousands. ACh is the number of authentic-data-trained checkpoints in an average of 8 checkpoints.

3.3 Translation Process

We carried out the translation in two stages: MT post-editing stage and final revision stage. For both stages, we used Memsource as the computer-assisted translation (CAT) tool.

(1) Post-editing The documents were first translated by all 13 MT systems. In addition, we included a variant with no translation ("Source") and with a pre-existing reference translation ("Reference").[8] The translated files were shuffled at document boundaries so that each document in the file was translated by a single MT system and no MT system appeared twice in a single file.

The resulting 15 files were given to 15 professional post-editors. Every post-editor worked with all 99 lines. This stage provides the primary data for determining the amount of time the post-editors need to bring the candidates to the common industry standards. The post-editors are well used to carrying out this task and are familiar with the CAT tool. In the translation editor, the MT

[4] cloud.google.com/translate

[5] azure.microsoft.com/en-us/services/cognitive-services/translator/

[6] We also experimented with BERTScore but its Pearson correlation with BLEU is 0.9939. This would lead to the same observations and conclusions.

[7] TER+t.tercom-nonorm-punct-noasian-uncased+v.1.5.1 BLEU+c.mixed+#.1+s.exp+tok.13a+v.1.4.14

[8] For simplicity, we refer to these two types of input (Reference and Source) also as MT systems.

outputs (incl. Reference, indistinguishable) were offered as 100% TM matches. No other technical tools (MT, TM etc.) were allowed. The post-editors received instructions mentioning that the provided translation may be manual or automatic (or missing in the case of Source). They were also asked to take any necessary breaks only at document boundaries marked in the input file.

(2) Revision After the first post-editing, the results were examined by 17 professional reviewers and further refined. None of the first-phase post-editors was included in the set of reviewers. Before submitting the data for the second stage, we further shuffled the translations on the document-level so that each reviewer received a random mix of documents produced by different post-editors.

In addition to the post-edited documents, we also included the pre-existing reference translation and the output of the document-level MT system (M11) without post-editing.

Again, "revision" is a standard task in the industry. The proposed translation is pre-filled in the output fields and reviewers modify it as necessary. The instructions mentioned that the proposed translation may be the result of manual post-editing of MT, manual translation with the help of MT or unedited MT output but suggested to fix only true errors: wrong translation, inaccuracies, grammar or style errors. Individual translators' preferences were supposed to be avoided.

The main goal of this stage is to measure how the quality of MT for post-editing affects the quality of the final translations. The standard practice of the industry is to focus on the output, providing the final quality check and the last necessary fixes, not any laborious error annotation. To complement purely quantitative measures (review time and string difference of the translation before and after the review), we asked the reviewers to also perform a basic variant of LQA ("Linguistic Quality Assurance") based on the MQM-DQF framework.[9] The reviewers classified errors into three categories: accuracy (adequacy), fluency and other. Every comment also had a severity attached (0-neutral, 1-minor, 2-major, 3-critical). We report severity averaged per number of sentences.

3.4 Post-Editors

We examine in detail the composition of our set of post-editors from the first phase, not of the reviewers. The experience in translating and post-editing of our post-editors, shown in Table 2, was slightly lower than that of Sanchez-Torron and Koehn (2016).

(years)	< 5	< 10	≥ 10
Translation experience	5	2	9
Post-Editor experience	8	7	0

Table 2: Post-editors' (first phase) years of translation and post-editing experience.

Questionnaire All post-editors were asked to fill in a questionnaire, which is inspired by Sanchez-Torron and Koehn (2016) but further extends it with questions regarding translators' sentiment on MT. The results are shown in Table 3 and are comparable to the data collected by Sanchez-Torron and Koehn (2016) with the exception of a slightly higher positive opinion of using MT. Namely, we see a higher agreement on *MT helps to translate faster* and less disagreement on *Prefer PE to editing human translation*. For questionnaire results of reviewers please see Appendix A.

There is a clear preference for using even imprecise TM matches (85–94%) over MT output. This corresponds to the general tendency towards the opinion that post-editing is more laborious than using a TM, which is an interesting contrast to the preference for post-editing over translation from scratch. The question about preferring to post-edit human over machine output shows a perfect Gaussian distribution, i.e. the lack of such preference in general. We see some level of trust in MT in the process helping to improve translation consistency and produce overall better results. For personal use, the post-editors know MT technology and use it for languages they do not speak.

4 Post-Editing Effort

To measure the post-editing effort, we examine first the differences between provided translation and the post-edited output (Section 4.1). We then focus on the time spent post-editing, which is an extrinsic evaluation of this task (Section 4.2).

4.1 Edits

Output Similarity The post-edited outputs of MT systems had 21.77 ± 0.11 tokens per line, which is slightly higher than for the original candidates (21.10 ± 0.12). Also, Reference (21.76

[9] qt21.eu/wp-content/uploads/2015/11/QT21-D3-1.pdf

Question	Response
Comfortable with post-editing human-like (perfect) quality	
Comfortable with post-editing less-than-perfect quality	
Prefer PE to translating from scratch (without a TM)	
MT helps to maintain translation consistency	
MT helps to translate faster	
PE is more laborious than translating from scratch or with a TM	
Prefer PE to processing 85–94% TM matches	
Prefer PE to editing a human translation	
MT helps to produce better results	
Often use MT outside of work for known languages	
Often use MT outside of work for unknown languages	

Table 3: Post-editors' (first phase) answers regarding their profession on the Likert scale (leftmost bar = Strongly Disagree, rightmost bar = Strongly Agree), TM = translation memory.

Model	**P0→P1**	**P1→P2**	**P0→P2**
Source	0.23	0.88	0.23
M01	0.65	0.94	0.63
M02	0.75	0.92	0.71
M03	0.72	0.90	0.69
M04	0.74	0.88	0.70
M05	0.74	0.94	0.73
M06	0.77	0.93	0.74
M07	0.80	0.93	0.78
M08	0.77	0.94	0.76
M09	0.77	0.93	0.76
M10	0.77	0.94	0.77
M11	0.80	0.95	0.80
M11*	-	-	0.92
Google	0.80	0.93	0.76
Microsoft	0.74	0.91	0.70
Reference	0.90	0.96	0.87
Reference*	-	-	0.87
Average	0.73	0.93	0.73
Lin. fit, all	0.011	0.001	0.015
Lin. fit, >36	0.004	0.000	0.027

Table 4: Average ChrF similarity per system between different stages of post-editing. Bottom two lines show linear fit coefficient on either all MT systems or on MT systems with BLEU > 36 (reference and source excluded). P0: system output, P1: post-editors' output, P2: reviewers' output.

tokens per line) got in comparison to MT systems less long in post-editing, reaching 22.10.[10]

To measure the distance between the provided translations and the post-editors' output, we used character n-gram F-score (ChrF, Popović, 2015). For computation we again use SacreBLEU[11] (Post, 2018).

Table 4 shows the measured ChrF similarities. For Source, the English input was used and received a similarity of 0.23 (caused by named entities, numbers and punctuation which can remain unchanged). On the opposite end, Reference had the highest similarity followed by Google, M11 and M07. The last two columns and linear fits are discussed in Section 4.3.

The post-editing of Reference had on average ChrF of 0.90, while MT models 0.75 ± 0.04. An improvement in one BLEU point then corresponds to 1.47% of the MT average ChrF.

Figure 1 shows the trend of the relationship of MT quality and post-edited distance (first phase, P0→P1) measured by ChrF2. It is systematically positive even when considering only top-n MT systems. Graphs for P1→P2 and P0→P2 (not shown) suggest that there is also a small correlation between the MT systems' quality and post-edited distance for the second phase (P1→P2). This trend is similar when using TER instead of BLEU and the figure is shown in Appendix B.

Unigram Comparison To examine the proportion of words that were only moved within the sentence, we also computed unigram precision and recall between the provided translations and the post-edited outputs. As expected, most of the words in the reference translation were unchanged by the post-editors ($F_1 = 0.92$), this is in contrast to the MT systems ($F_1 = 0.78 \pm 0.04$).

In this case, the linear relationship to BLEU is

[10] 95% confidence interval based on Student's t-test.

[11] ChrF2+numchars.6+space.false+version.1.4.14

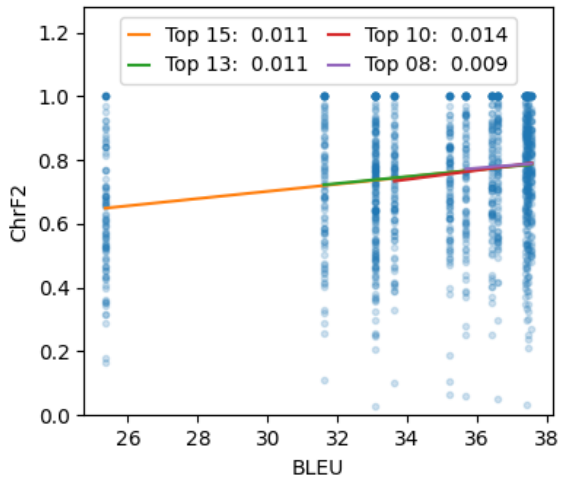

Figure 1: Sentence similarity measured by ChrF2 between the provided translation and first-phase (P0→P1). Every dot is a single sentence translated by a given MT. Source and Reference measurements are omitted for scale.

preserved from all models (slope 0.011) to only the top eight models (slope 0.008). The first corresponds to 1.41%.

4.2 Time

We focus on time spent per one token, which is more useful in determining the overall time a post-editor has to spend working with a document than time spent per one sentence. The CAT tool records two quantities for each segment:

- Think time: the time between entering a given segment (i.e. the keyboard cursor moves to that segment) and doing the first edit operation
- Edit time: the sum of thinking time and the time spent editing the segment (until the translation is finished and confirmed)

However, both of these measured quantities carry some level of noise due to translator breaks and other distractions in the think time.

The post-editors were instructed to take breaks only at document boundaries, but there were a number of deviations[12] which can be explained only by the post-editor getting distracted by other activities. Most of these deviations were present already in think time. Let T and W be the true variables for think and write times per word and $\hat{T}$ and $\hat{A}$ (all time) our measured estimates. The term ϵ_T is then causing the high deviation in $\hat{T}$ and subsequently in $\hat{A}$.

$$\begin{aligned}
&\hat{T} \approx T + \epsilon_T && \text{Measured think time}\\
&\hat{A} \approx \hat{T} + \hat{W} && \text{Measured total time}\\
&\quad = T + W + \epsilon_T + \epsilon_W &&\\
&\overset{*}{W} := \hat{A} - \hat{T} && \text{Measured write time}\\
&\quad \approx W + \epsilon_W &&\\
&\overset{*}{T} := \min\{10s, \hat{T}\} && \text{Estimated think time}\\
&\overset{*}{A} := \overset{*}{W} + \min\{10s, \hat{T}\} && \text{Estimated total time}
\end{aligned}$$

The two quantities $\overset{*}{T}$ and $\overset{*}{A}$ then approximate the think time and total time, respectively. The latter is used in the following figures and referred to as estimated total time. Think time is estimated by capping the value per word to 10s.[13] The choice of filtering has a significant impact on the result. Even though the current strategy was chosen with the best intentions, it is unclear whether it is universally the optimal one. The interest in the variable of the total time is sparked by the immediate commercial relevance: *How does one BLEU point in used MT system affect the total work time of post-editors spent on one word?*[14]

Table 5 shows the estimates for think and total times with 95% confidence intervals. Although there is some overlap between the systems, they are spread out evenly between 6s and 12s. Reference ranked by far the first and Source in the middle. For all systems below Source, post-editing MT output took longer than translating from scratch.

The relationship between BLEU and total time per word is shown in Figure 2. It shows no clear systematic relationship between the two variables. For comparison with Sanchez-Torron and Koehn (2016) we also report slopes of linear least-squares fits. A slope of 0.044 (all MT systems) indicates that a 1 BLEU point increase in MT quality increases total time per word by 0.044s (i.e., that higher-quality MT may in fact lead to slightly longer post-editing). For top-8 systems, the slope is negative, meaning that a 1 BLEU point increase

[12] Maximum time per word was 1482s, which is highly improbable.

[13] We chose 10 seconds to cap the think time per word because it seemed improbable that anyone would genuinely spend all this time thinking about the upcoming sentence.

[14] Prices of translations are usually calculated by the number of words in the document.

Model	Total time	Think time
Reference	3.17s±0.13s	0.58s±0.04s
M08	4.10s±0.20s	0.55s±0.03s
Google	4.52s±0.22s	0.96s±0.08s
M03	4.60s±0.19s	0.60s±0.04s
M07	4.95s±0.27s	0.92s±0.06s
M01	5.13s±0.18s	0.97s±0.05s
M09	5.41s±0.36s	1.12s±0.07s
M05	5.64s±0.21s	0.93s±0.07s
Source	6.00s±0.22s	0.72s±0.05s
Microsoft	6.02s±0.32s	0.87s±0.06s
M04	6.27s±0.27s	1.46s±0.09s
M02	6.44s±0.27s	1.16s±0.07s
M10	6.45s±0.32s	2.31s±0.12s
M11	8.01s±0.47s	1.63s±0.09s
M06	8.25s±0.39s	1.62s±0.07s
Average	5.66s±0.07s	1.09s±0.02s

Table 5: Total and think time estimations for first phase of post-editing for all MT systems (+Source and Reference). Confidence intervals computed for 95%. Sorted by total time.

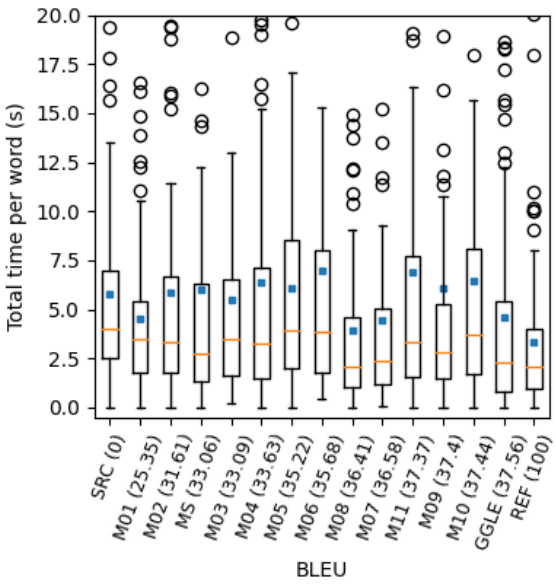

Figure 2: Total time per word in relation to MT system BLEU score. Every dot is a single post-edited sentence. Zoomed to [0, 20] on the y-axis. Orange bars represent medians and blue squares means. Upper whiskers are the 3rd quartile + 1.5× inter-quartile range.

decreases total word time by 0.514s. However, these results should not be interpreted in the sense that for high-quality MT systems, BLEU improvements lead to faster post-editing. On the contrary, they should illustrate the uncertainty and complexity of the relationship between the two quantities.

4.3 Quality

The quality of post-editors' output was measured during revision. This closely follows industry standards, where the text to translate is first given to post-editors and then to another set of reviewers. Here we again used ChrF to determine how much effort was needed to create production-level translations from the already post-edited translations.

Apart from the outputs of the work of the first phase of post-editors, we also mixed in unedited Reference (labelled Reference*) and M11 (M11*). This allows us to see if there is any effect of priming the workers with the task specification: post-editors are likely to expect to have more work with fixing MT output than reviewers.

In this case, the total and think times were estimated the same way as for the first phase. The results per system are shown in Table 6. The distribution is now more uniform, and in comparison to the first phase, shown in Table 5, many systems changed their rank. Documents of M11* and Reference* (not post-edited in the first phase) had much larger average total times than their post-edited versions, M11 and Reference. This is caused by more thorough reviewing necessary since the documents were not refined. Furthermore, the reviewers may not have expected this kind of input at all. Note however that the total time for M11* is still not much higher than the average time required to review an already post-edited MT output.

In contrast to Figure 2, the linear least-square fit slope for total times of top-15 and top-8 are 0.069 and 0.765 in the case of reviewing. This suggests that an improvement in BLEU may lead to higher times when reviewing the post-edited output and that BLEU may be not a good predictor of overall localization effort. We currently do not have an explanation for this effect.

The reviewers were also tasked to mark errors using LQA. For every sentence, we sum the LQA severities and compute an average for every model. There was no significant linear relationship between the average severity and overall performance measured by BLEU. Exceptions (deviating from the average 0.51) are Reference* (0.59) and M11* (0.83), which were not post-edited in

Model	Total time	Think time
M08	2.12s±0.11s	0.96s±0.07s
M01	2.29s±0.14s	0.96s±0.06s
Reference	2.32s±0.12s	0.97s±0.06s
M11	2.34s±0.11s	1.10s±0.06s
M06	2.53s±0.17s	0.96s±0.05s
M02	2.98s±0.18s	0.83s±0.04s
Google	3.12s±0.13s	1.31s±0.07s
M07	3.36s±0.22s	1.19s±0.08s
Source	3.37s±0.12s	1.01s±0.05s
M04	3.70s±0.13s	1.10s±0.06s
M05	3.75s±0.28s	1.05s±0.06s
Microsoft	3.75s±0.22s	1.12s±0.06s
M11*	3.96s±0.30s	1.17s±0.08s
M03	4.06s±0.16s	0.87s±0.05s
M09	4.41s±0.23s	0.85s±0.06s
M10	4.83s±0.31s	1.71s±0.08s
Reference*	5.31s±0.18s	1.52s±0.07s
Average	3.42s±0.05s	1.10s±0.02s

Table 6: Total and think time estimations for the review phase of post-editing for all MT systems (+Source and Reference). Confidence intervals computed for 95%. Sorted by total time.

the first phase. Source had an average severity of 0.61, while the best system, M11, had the lowest 0.28. There was, however, a significant difference between the average severity of documents: Lease (0.26), Audit (0.40), Technical (0.40) and News (0.74). The average LQA severity is shown in Table 7.

Output Similarity Table 4 shows the similarities between the output of the first phase and the second phase (second column) and the system output and the second phase (third column). For M11* and Reference*, the output of the first phase is undefined. The similarities in the second column are much more dispersed, though still Reference was post-edited the less (0.96, while Source the most (0.88). A similar thing can be observed between the system output and final output, with the exception of non-post-edited M11* being post-edited very little. In fact, the raw MT output was modified less than some of the already post-edited translations. Reference post-edited by first phase post-editors had a similarity to the original 0.90, which is very similar to Reference post-edited only by the reviewer (Reference*, 0.87).

Linear functions are fitted to see the effect of

Model/Doc.	Acc.	Flu.	Other	All
Source				
M01				
M02				
M03				
M04				
M05				
M06				
M07				
M08				
M09				
M10				
M11				
M11*				
Google				
Microsoft				
Reference				
Reference*				
News				
Audit				
Technical				
Lease				

Table 7: Average LQA severity (reported from 0 to 3) of models and documents across three categories: Adequacy/accuracy, fluency and other. Their average is reported in the last column. Empty and full squares represent severities of 0 and 1, respectively.

one BLEU point on the amount of post-editing (measured by ChrF). The results are in the bottommost lines of Table 4. The effect is the strongest when measuring the similarity between the model output and the second phase. The linear fit is, however, strongly influenced by the less-performing models. In the case where only the top eight models (BLEU $>$ 36) are taken, an increase of 1 BLEU point corresponds to 0.027 increase in similarity between model output and final version of the sentence ($\sim 3.7\%$ of the total average). A similar trend (negative slope) was observed also when using TER instead of BLEU. Reference and Source were excluded from this computation because their artificial BLEU scores (100 and 0 respectively) would have an undesired effect on the

result. The average similarity between the once post-edited output and the corrections is 0.93, confirming the hypothesis that most of the errors are resolved in the first pass of post-editing.

Editing process. Table 8 shows the breakdown of edit types[15] between the provided translation and the post-edited output. The *insert* and *delete* values show no apparent trend, though *replace* is decreasing with higher BLEU scores.

Model	Replace	Delete	Insert
Source	22.22	0.00	0.02
M01	9.71	0.64	0.69
M02	6.84	0.30	0.66
M03	7.48	0.51	0.71
M04	6.49	0.41	0.54
M05	6.67	0.48	0.40
M06	6.32	0.78	0.59
M07	5.49	0.33	0.62
M08	6.03	0.46	0.38
M09	5.84	0.32	0.52
M10	5.70	0.61	0.68
M11	5.62	0.40	0.94
Google	5.32	0.32	0.44
Microsoft	7.29	0.42	0.98
Reference	2.96	0.18	0.42
Average	7.34	0.41	0.57

Table 8: Average number of line edit operations for the first phase of post-editing for all MT systems (+Source and Reference). For specific operations, **Insert** considers the number of target tokens, **Delete** the number of source tokens and **Replace** their average.

5 Summary

In this work, we extended the standard scenario for testing post-editing productivity by a second phase of annotations. This allowed for further insight into the quality of the output and it also follows the standard two-phase process of translation companies more closely.

We found a complex relationship between MT quality and post-editing speed, which depends on many factors. When considering only the top 8 systems, an improvement of one BLEU point corresponded to 0.514 fewer seconds per one word on average but at the same time, this trend was not confirmed on larger sets of systems. Overall, the relationship is most likely weaker than previously assumed.

We did not find any significant relationship between the produced output quality and MT system performance among all systems because the effect was not measurable in the second phase. As expected, post-editing human reference led to the smallest amount of edits and time spent. Contrary to current results, translating from scratch was not significantly slower than post-editing in either of the two phases. The average ChrF similarity between the provided output and the first phase results was 0.73 and between the two phases 0.93, suggesting diminishing results of additional phases.

The most significant conclusion is that for NMT, the previously assumed link between MT quality and post-editing time is weak and not straightforward. The current recommendation for the industry is that they should not expect small improvements in MT (measured by automatic metrics) to lead to significantly lower post-editing times nor significantly higher post-edited quality.

Ethics

Both post-editors and proofreaders were compensated by their usual professional wages.

Acknowledgments

We sincerely thank České překlady for their collaboration and for providing all translations and revisions. The work was supported by Memsource and by the grants 19-26934X (NEUREM3) and 20-16819X (LUSyD) by the Czech Science Foundation.

The work has been using language resources developed and distributed by the LINDAT/CLARIAHCZ project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2018101).

References

Eneko Agirre, António Branco, Martin Popel, and Kiril Simov. 2015. QTLeap WSD/NED corpus. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Sheila Castilho, Joss Moorkens, Federico Gaspari, Iacer Calixto, John Tinsley, and Andy Way. 2017.

[15] Using the Ratcliff-Obershelp algorithm (Ratcliff and Metzener, 1988) implemented in Python `difflib`.

Is neural machine translation the new state of the art? *The Prague Bulletin of Mathematical Linguistics*, (108).

Philipp Koehn and Ulrich Germann. 2014. The impact of machine translation quality on human post-editing. In *Proceedings of the EACL 2014 Workshop on Humans and Computer-assisted Translation*, pages 38–46.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. *arXiv preprint arXiv:1706.03872*.

Maarit Koponen. 2013. This translation is not too bad: An analysis of post-editor choices in a machine translation post-editing task. In *Workshop Proceeding: Workshop on Post-editing Technology and Practice (WPTP-2)*, pages 1–9.

Maarit Koponen. 2016. Is machine translation post-editing worth the effort? a survey of research into post-editing and effort. *The Journal of Specialised Translation*, 25:131–148.

Maarit Koponen and Leena Salmi. 2015. On the correctness of machine translation: A machine translation post-editing task. *Journal of Specialised Translation*, 23:118–136.

Maarit Koponen, Leena Salmi, and Markku Nikulin. 2019. A product and process analysis of post-editor corrections on neural, statistical and rule-based machine translation output. *Machine Translation*, 33(1):61–90.

Maarit Koponen, Umut Sulubacak, Kaisa Vitikainen, and Jörg Tiedemann. 2020. MT for subtitling: User evaluation of post-editing productivity. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 115–124.

Samuel Läubli, Chantal Amrhein, Patrick Düggelin, Beatriz Gonzalez, Alena Zwahlen, and Martin Volk. 2019. Post-editing productivity with neural machine translation: An empirical assessment of speed and quality in the banking and finance domain. *arXiv preprint arXiv:1906.01685*.

Linda Mitchell, Johann Roturier, and Sharon O'Brien. 2013. Community-based post-editing of machine-translated content: monolingual vs. bilingual. In *Proceedings of the MT Summit Conference 2013*. European Association for Machine Translation.

Mirko Plitt and François Masselot. 2010. A productivity test of statistical machine translation post-editing in a typical localisation context. *The Prague bulletin of mathematical linguistics*, 93(1):7–16.

Martin Popel. 2020. CUNI English-Czech and English-Polish Systems in WMT20: Robust Document-Level Training. In *Proceedings of the Fifth Conference on Machine Translation*, pages 269–273, Online. Association for Computational Linguistics.

Martin Popel, Marketa Tomkova, Jakub Tomek, Łukasz Kaiser, Jakob Uszkoreit, Ondřej Bojar, and Zdeněk Žabokrtský. 2020. Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals. *Nature Communications*, 11(4381):1–15.

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

John W Ratcliff and David E Metzener. 1988. Pattern-matching-the gestalt approach. *Dr Dobbs Journal*, 13(7):46.

Marina Sanchez-Torron and Philipp Koehn. 2016. Machine translation quality and post-editor productivity. In *Proceedings of the Conference of the Association for Machine Translation in the Americas (AMTA) Vol. 1: MT Researchers' Track*, pages 16–26. Association for Machine Translation in the Americas, AMTA. Twelfth Conference of The Association for Machine Translation in the Americas, AMTA 2016 ; Conference date: 28-10-2016 Through 01-11-2016.

Germán Sanchis-Trilles, Vicent Alabau, Christian Buck, Michael Carl, Francisco Casacuberta, Mercedes García-Martínez, Ulrich Germann, Jesús González-Rubio, Robin L Hill, Philipp Koehn, et al. 2014. Interactive translation prediction versus conventional post-editing in practice: a study with the casmacat workbench. *Machine Translation*, 28(3-4):217–235.

Midori Tatsumi. 2009. Correlation between automatic evaluation metric scores, post-editing speed, and some other factors. *The Twelfth Machine Translation Summit (MT-Summit XII)*, pages 332–339.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Lrec*, volume 2012, pages 2214–2218.

Jörg Tiedemann, Tommi Nieminen, Mikko Aulamo, Jenna Kanerva, Akseli Leino, Filip Ginter, and Niko Papula. 2020. The fiskmö project: Resources and tools for finnish-swedish machine translation and cross-linguistic research. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 3808–3815.

Vilém Zouhar, Tereza Vojtěchová, and Ondřej Bojar. 2020. WMT20 document-level markable error exploration. In *Proceedings of the Fifth Conference on Machine Translation*, pages 371–380.

A Questionare for Second-Phase

(years)	< 5	< 10	≥ 10
Translation experience	1	2	14
Post-Editor experience	10	3	4

Table 9: Reviewers' (second phase) years of translation and post-editing experience.

Question	Response
Comfortable with post-editing human-like (perfect) quality	
Comfortable with post-editing less-than-perfect quality	
Prefer PE to translating from scratch (without a TM)	
MT helps to maintain translation consistency	
MT helps to translate faster	
PE is more laborious than translating from scratch or with a TM	
Prefer PE to processing 85–94% TM matches	
Prefer PE to editing a human translation	
MT helps to produce better results	
Often use MT outside of work for known languages	
Often use MT outside of work for unknown languages	

Table 10: Reviewers' (second phase) answers regarding their profession on the Likert scale (leftmost bar = Strongly Disagree, rightmost bar = Strongly Agree), TM = translation memory.

B TER as an Evaluation Measure

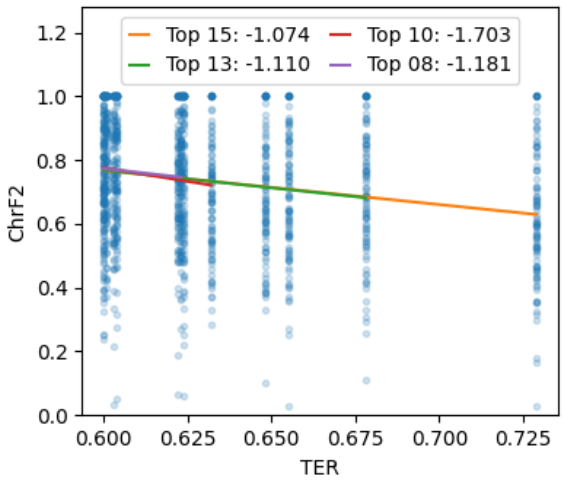

Figure 3: Sentence similarity measured by ChrF2 between the provided translation and first-phase (P0→P1) in contrast to system TER score (lower is better). Every dot is a single sentence translated by a given MT. Source and Reference measurements are omitted for scale.