DivEMT: Neural Machine Translation Post-Editing Effort Across Typologically Diverse Languages

Gabriele Sarti, **Arianna Bisazza**, **Ana Guerberof-Arenas**, **Antonio Toral**

Center for Language and Cognition (CLCG), University of Groningen
{g.sarti, a.bisazza, a.guerberof.arenas, a.toral.ruiz}@rug.nl

Abstract

We introduce DivEMT, the first publicly available post-editing study of Neural Machine Translation (NMT) over a typologically diverse set of target languages. Using a strictly controlled setup, 18 professional translators were instructed to translate or post-edit the same set of English documents into Arabic, Dutch, Italian, Turkish, Ukrainian, and Vietnamese. During the process, their edits, keystrokes, editing times and pauses were recorded, enabling an in-depth, cross-lingual evaluation of NMT quality and post-editing effectiveness. Using this new dataset, we assess the impact of two state-of-the-art NMT systems, Google Translate and the multilingual mBART-50 model, on translation productivity. We find that post-editing is consistently faster than translation from scratch. However, the magnitude of productivity gains varies widely across systems and languages, highlighting major disparities in post-editing effectiveness for languages at different degrees of typological relatedness to English, even when controlling for system architecture and training data size. We publicly release the complete dataset[1] including all collected behavioral data, to foster new research on the translation capabilities of NMT systems for typologically diverse languages.

1 Introduction

Recent advances in neural language modeling and multilingual training have prompted a widespread adoption of machine translation (MT) technologies across an unprecedented range of world languages. While the benefits of state-of-the-art MT for cross-lingual information access are undisputed (Lommel and Pielmeier, 2021), its usefulness as an aid to professional translators varies considerably across domains, subjects and language combinations (Zouhar et al., 2021). In the last decade, the MT community has been including an increasing number of languages in its automatic and human evaluation efforts (Bojar et al., 2013; Barrault et al., 2021). However, the results of these evaluations are typically not directly comparable across different language pairs for various reasons. First, reference-based automatic quality metrics are hardly comparable across different target languages (Bugliarello et al., 2020). Secondly, human judgments are collected independently for different language pairs, making their cross-lingual comparison vulnerable to confounding factors such as tested domains and training data sizes. Similarly, recent work on NMT post-editing efficiency has focused on specific language pairs such as English-Czech (Zouhar et al., 2021), German-Italian, German-French (Läubli et al., 2019) and English-Hindi (Ahsan et al., 2021), but a controlled comparison across a set of typologically diverse languages is still lacking.

In this work, we assess the usefulness of state-of-the-art NMT in professional translation with a strictly controlled cross-language setup (Figure 1). Specifically, professionals were asked to translate the same English documents into six typologically different languages (Arabic, Dutch, Italian, Turkish, Ukrainian, and Vietnamese) using the same platform and guidelines. Three *translation modalities* were adopted: human translation from scratch (HT), post-editing of Google Translate's translation (PE_1), and post-editing of mBART-50's translation (PE_2), the latter being a state-of-the-art open-source, multilingual NMT system. In addition to post-editing results, subjects' fine-grained editing behavior — including keystrokes and time information — was logged to measure productivity and effort across languages, systems and translation modalities. Finally, translators were asked to complete a qualitative assessment regarding their perceptions of MT quality and post-editing effort. The resulting DivEMT dataset is to our best knowledge the first public resource allow-

[1] https://github.com/gsarti/divemt
https://huggingface.co/datasets/GroNLP/divemt

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 7795–7816
December 7-11, 2022

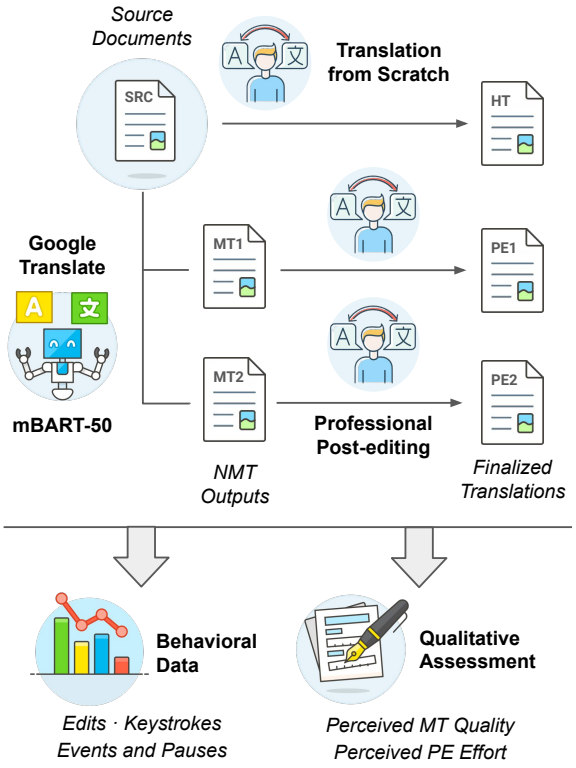

Figure 1: The DivEMT data collection process. For every English source document, 18 professional translators are tasked to translate it from scratch (HT) or post-edit NMT systems' outputs (PE_1/PE_2) into six typologically diverse target languages. Behavioral data and qualitative assessments are collected during and after the process respectively.

ing a direct comparison of professional translators' productivity and fine-grained editing information across a set of typologically-diverse languages. DivEMT is publicly released alongside this paper as a unique resource to study the language- and system-dependent nature of NMT advances in real-world translation scenarios.

2 Related Work

Cross-lingual MT Evaluation Before the advent of NMT, Birch et al. (2008) studied how various language properties affected the quality of Statistical MT (SMT) across a sizeable sample of European language pairs. The comparison, however, was solely based on BLEU, which is in fact not comparable across different target languages (Bugliarello et al., 2020). Recent work on neural models introduced more principled ways to measure the intrinsic difficulty of language-modeling (Gerz et al., 2018; Cotterell et al., 2018; Mielke et al., 2019) and machine-translating (Bugliarello et al., 2020; Bisazza et al., 2021) different languages. However, achieving this reliably without any human evaluation remains an open research question. Human evaluations of MT quality are routinely conducted during campaigns such as WMT (Koehn and Monz, 2006; Akhbardeh et al., 2021) and IWSLT (Cettolo et al., 2016, 2017) among others, but their focus is on language- and domain-specific ranking of MT systems — often leveraging non-professional annotators (Freitag et al., 2021) — rather than cross-lingual quality comparisons. Concurrently to this work, Licht et al. (2022) proposed a new human evaluation protocol to improve consistency in cross-lingual MT quality assessment.

Post-editing NMT Measuring post-editing effort across its *temporal, cognitive, and technical* dimensions (Krings, 2001) is a well-established way to assess the effectiveness and efficiency of MT as a component of specialized translation workflows. Seminal post-editing studies highlighted an increase in translators' productivity following MT adoption (Guerberof, 2009; Green et al., 2013; Läubli et al., 2013; Plitt and Masselot, 2010; Parra Escartín and Arcedillo, 2015). However, they also struggled to identify generalizable findings due to confounding factors like output quality, content domains, and high variance across language pairs and human subjects. With the advent of NMT, productivity gains of the new approach were extensively compared to those of SMT, the highly-customized dominant paradigm at the time (Castilho et al., 2017; Bentivogli et al., 2016; Toral et al., 2018; Läubli et al., 2019). Initial results were promising for NMT due to its better fluency and overall results. Moreover, translators were shown to prefer NMT over SMT for post-editing, although a pronounced productivity increase was not always present. More recent work highlighted the productivity gains driven by NMT post-editing in a wider array of languages that were previously challenging for MT, such as English-Dutch (Daems et al., 2017), English-Hindi (Ahsan et al., 2021), English-Greek (Stasimioti and Sosoni, 2020), English-Finnish and English-Swedish (Koponen et al., 2020), all showing a considerable variance among language pairs and subjects. Interestingly, Zouhar et al. (2021) found NMT post-editing speed to be comparable to translation from scratch in English-Czech, and highlighted a disconnect between moderate increases in automatic MT

quality metrics and better post-editing productivity. In sum, research on post-editing NMT generally reports increased fluency and output quality, but productivity gains are hardly generalizable across language pairs and domains. Importantly, to our knowledge, no previous work has studied NMT post-editing over a set of typologically different languages while controlling for the effects of content types and domains, NMT engines, and translation interfaces.

3 The DivEMT Dataset

DivEMT's main purpose is to assess the usefulness of state-of-the-art NMT for professional translators and to study how this usefulness varies across target languages with different typological properties. We present below our data collection setup, which strikes a balance between simulating a realistic professional translation workflow and maximizing the comparability of results across languages.

3.1 Subjects and Task Scheduling

To control for the effect of individual translators' preferences and styles, we involve a total of 18 subjects (three per target language). During the experiment, each subject receives a series of short *documents* (3 to 5 sentences each) where the source text is presented in isolation (HT) or alongside a translation proposal produced by one of the NMT systems (PE_1, PE_2). The experiment comprises two phases: During the **warm-up phase** a set of 5 documents is translated by all subjects following the same, randomly sampled sequence of modalities (HT, PE_1 or PE_2). This phase allows the subjects to get used to the setup and enables us to spot possible issues in the logged behavioral data before moving forward.[2] In the **main collection phase**, each subject is asked to translate documents in a pseudo-random sequence of modalities. This time, however, the sequence is different for each translator and chosen so that each document gets translated in all three modalities. This allows us to measure translation productivity independently from the subject's productivity and document-specific difficulties. A graphical overview of this process is shown in Figure 1, with additional details given in Appendix A. As productivity and other behavioral metrics can only be estimated with a sizable sample, we prioritize the number of documents over the number of subjects per language during budget allocation. A larger set of post-edited documents also provides more insight in the error type distribution of NMT systems across different language pairs, an analysis which we leave to future work.

All subjects are professional translators with at least 3 years of professional experience, at least one year of post-editing experience and strong proficiency with CAT tools.[3] Translators were provided with links to the source articles to facilitate contextualization, were asked to produce translations of publishable quality and were instructed not to use any external MT engine to produce their translations. Assessing the final quality of the post-edited material is out of the scope of the current study, although we realize that this is an important consideration to assess usability in a professional context. A summary of our translation guidelines is provided in Appendix C.

3.2 Choice of Source Texts

The selected documents represent a subset of the FLORES-101 benchmark (Goyal et al., 2022) consisting of sentences taken from English Wikipedia, and covering a mix of topics and domains.[4] While professional translators generally specialize in one or a few domains, we opt for a mix-domain dataset to minimize domain adaptation efforts by the subjects and maximize the generalizability of our results. Importantly, FLORES-101 includes high-quality human translations into 101 languages, which makes it possible to automatically estimate NMT quality and discard excessively low-scoring models or language pairs before our experiment. FLORES-101 also provides useful metadata, e.g. source URL, which allows us to ensure the absence of public translations of the selected contents, which could be leveraged by translators and compromise the validity of our setup. The documents used for our study are fragments of contiguous sentences extracted from Wikipedia articles that compose the original FLORES-101 corpus. Even if small, the context provided by document structure allows us to simulate a more realistic translation workflow if compared to out-of-context sentences.

Based on our available budget, we select 112 English documents from the *devtest* portion of FLORES-101 corresponding to 450 sentences and 9626 words. More details on the data selection process are provided in Appendix D.

[2] Warm-up data are excluded from the analysis of Section 4.

[3] Additional subjects' details are available in Appendix B.

[4] We use a balanced sample of articles sourced from WikiNews, WikiVoyage and WikiBooks.

	Genus:Family	d_{syn}	Morph.	MSP	TTR	Script
ENG	IE:Germanic	–	Fus	1.17	0.28	latin
ARA	Af:Semitic	0.57	Ifx	1.67	0.46	arab.
NLD	IE:Germanic	0.49	Fus	1.16	0.28	latin
ITA	IE:Romance	0.51	Fus	1.30	0.30	latin
TUR	Alt:Turkic	0.70	Agg	2.28	0.50	latin
UKR	IE:Slavic	0.51	Fus	1.42	0.47	cyril.
VIE	Au:VietMuong	0.57	Iso	1.00	0.12	latin

Table 1: Typological diversity of our language sample. **IE**: Indo-European, **Af**: Afro-Asiatic, **Alt**: Altaic, **Au**: Austro-Asiatic. d_{syn}: Syntactic distance w.r.t. English (Lin et al., 2019). **Fus**: fusional, **Ifx**: introflexive, **Agg**: agglutinative, **Iso**: isolating. **MSP**: Mean size of paradigm, from Çöltekin and Rama (2022). **TTR**: Type-token ratio measured on FLORES-101. Shading indicates genetic/syntactic relatedness to English and morphological complexity/lexical richness.

3.3 Choice of Languages

Training data is among the most important factors in defining the quality of a NMT system. Unfortunately, using strictly comparable or multi-parallel datasets, like Europarl (Koehn, 2005) or the Bible corpus (Mayer and Cysouw, 2014), would dramatically restrict the diversity of languages available to our study, or imply a prohibitively low translation quality on general-domain text. In order to minimize the effect of training data disparity while maximizing language diversity, we choose representatives of six different language families for which comparable amounts of training data are available in our open-source model, namely **Arabic**, **Dutch**, **Italian**, **Turkish**, **Ukrainian**, and **Vietnamese**. As shown in Table 1, our language sample ensures a good diversity in terms of language family and relatedness to English, type of morphological system, morphological complexity — measured by mean size of paradigm (MSP, Xanthos et al. 2011) — and script. We also report type-token ratio (TTR), the only language property that was found to correlate significantly with translation difficulty in a sample of European languages (Bugliarello et al., 2020). While the amount of language-specific parallel sentence pairs used for the multilingual fine-tuning of mBART-50 varies widely (4K$<N<$45M), all our selected language pairs fall within the 100K-250K range (mid-resourced, see Table 2), enabling a fair cross-lingual performance comparison.

3.4 Choice of MT Systems

While most of the best-performing general-domain NMT systems are commercial, experiments based on such systems are not replicable as their backends get silently updated over time. Moreover, without knowing the exact training specifics, we cannot attribute differences in the cross-lingual results to intrinsic language properties. We balance these observations by including two NMT systems in our study: **Google Translate** (GTrans)[5] as a representative of commercial quality, and **mBART-50 one-to-Many**[6] (Tang et al., 2021) as a representative of state-of-the-art open-source multilingual NMT technology. The original multilingual BART model (Liu et al., 2020) is an encoder-decoder transformer model pre-trained on monolingual documents in 25 languages. Tang et al. (2021) extend mBART by further pre-training on 25 new languages and performing *multilingual translation fine-tuning* for the full set of 50 languages, producing three configurations of multilingual NMT models: many-to-one, one-to-many, and many-to-many. Our choice of mBART-50 is largely motivated by its maneageable size, its good performances across the set of evaluated languages (see Table 2) and its adoption for other NMT (Liu et al., 2021) and post-editing (Fomicheva et al., 2020) studies. Although mBART-50 performances are usually comparable or slightly worse than the ones of tested bilingual NMT models,[7] using a multilingual model allows us to evaluate the downstream effectiveness of a single, unified system trained on pairs evenly distributed across tested languages. Finally, adopting two systems with marked differences in automatic evaluation scores allows us to estimate how a significant increase in metrics such as BLEU, CHRF and COMET (Papineni et al., 2002; Popović, 2015;

	GTrans (PE_1)	mBART-50 (PE_2)	# Pairs
ARA	**34.1 / 65.6 / .737**	17.0 / 48.5 / .452	226K
NLD	**29.1 / 60.0 / .667**	22.6 / 53.9 / .532	226K
ITA	**32.8 / 61.4 / .781**	24.4 / 54.7 / .648	233K
TUR	**35.0 / 65.5 / 1.00**	18.8 / 52.7 / .755	204K
UKR	**31.1 / 59.8 / .758**	21.9 / 50.7 / .587	104K
VIE	**45.1 / 61.9 / .724**	34.7 / 54.0 / .608	127K

Table 2: MT quality of the selected NMT systems for English-to-Target translation on the full FLORES-101 devtest split, in BLEU / CHRF / COMET format. Best scores are highlighted in **bold**. We report the number of sentence pairs used for mBART-50 multilingual fine-tuning by Tang et al. (2021).

[5] Evaluation performed in October 2021.

[6] facebook/mbart-large-50-one-to-many

[7] See Appendix E for automatic MT quality results by five different models over a larger set of 10 target languages.

ENG	SRC	Inland waterways can be a good theme to base a holiday around.
ARA	HT	يمكن أن تكون الممرات المائية الداخلية خياراً جيدًا لتخطيط عطلة حولها.
	MT	يمكن أن تكون السكك الحديدية الداخلية موضوعًا جيدًا لإقامة عطلة حول.
	PE	قد تكون الممرات المائية الداخلية مكانًا جيدًا لقضاء عطلة حولها.
NLD	HT	Binnenlandse waterwegen kunnen een goed thema zijn voor een vakantie.
	MT	Binnenwaterwegen kunnen een goed thema zijn om een vakantie rond te zetten .
	PE	Binnenwaterwegen kunnen een goed thema zijn om een vakantie rond te organiseren .
ITA	HT	I corsi d'acqua dell'entroterra possono essere un ottimo punto di partenza da cui organizzare una vacanza.
	MT	I corsi d'acqua interni possono essere un buon tema per fondare una vacanza.
	PE	I corsi d'acqua interni possono essere un buon tema su cui basare una vacanza.
TUR	HT	İç bölgelerdeki su yolları, tatil planı için iyi bir tema olabilir.
	MT	İç suyolları, tatil için uygun bir tema olabilir.
	PE	İç sular tatil için uygun bir tema olabilir.
UKR	HT	Можна спланувати вихідні, взявши за основу подорож внутрішніми водними шляхами.
	MT	Водні шляхи можуть бути хорошим об 'єктом для базування відпочинку навколо .
	PE	Місцевість навколо внутрішніх водних шляхів може бути гарним вибором для організації відпочинку.
VIE	HT	Du lịch trên sông có thể là một lựa chọn phù hợp cho kỳ nghỉ.
	MT	Các tuyến nước nội địa có thể là một chủ đề tốt để xây dựng một kì nghỉ.
	PE	Du lịch bằng đường thủy nội địa là một ý tưởng nghỉ dưỡng không tồi.

Table 3: A DivEMT corpus entry, including the English source (SRC), its translation from scratch (HT), the MT output of mBART-50 (MT) and its post-edited version (PE) for all languages. We highlight insertions , deletions , substitutions and shifts computed with Tercom (Snover et al., 2006). Full examples available in Appendix F.

Rei et al., 2020) impacts downstream productivity across languages in a realistic post-editing scenario.

3.5 Translation Platform and Collected Data

Translators were asked to use PET (Aziz et al., 2012), a computer-assisted translation tool that supports both translating from scratch and post-editing. This tool was chosen because (i) it logs information about the post-editing process, which we use to assess effort (see Section 4); and (ii) it is a mature research-oriented tool that has been successfully used in several previous studies (Koponen et al., 2012; Toral et al., 2018). The minimalistic nature of PET interface and functionalities limits its application in commercial translation activities, making it generally unfamiliar for professional translators. We consider this aspect an advantage in light of our controlled setup since it allows us to avoid additional confounding effects or disparities stemming from tools-specific capabilities and different degrees of proficiency with the software. We also observe that, due to the varied and generic nature of the selected documents, functionalities such as concordance and translation memory matches would have proven much less useful in our setup. We collect three types of data:

- **Resulting translations** produced by translators in either HT or PE modes, constituting a multilingual corpus with one source text and 18 translations (one per language-modality combination) exemplified in Table 3.
- **Behavioral data** for translated sentences, including editing time, amount and type of keystrokes (content, navigation, erase, etc.), and number and duration of pauses above 300/1000 milliseconds (Lacruz et al., 2014).
- **Pre- and post-task questionnaire**. The former focuses on demographics, education, and work experience with translation and post-editing. The latter elicits subjective assessments of post-editing quality, effort and enjoyability compared to translating from scratch.

4 Post-Editing Effort Across Languages

In this section, we use the DivEMT dataset to quantify the post-editing effort of professional translators across our diverse set of target languages.

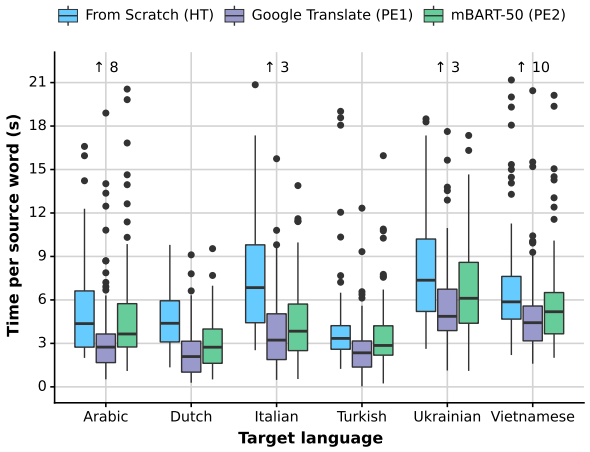

Figure 2: Temporal effort across languages and translation modalities, measured in seconds per processed source word. Each point represents a document, with higher scores denoting slower editing. ↑: amount of data points per language not shown in the plot.

We consider two main objective indicators of editing effort, namely *temporal measurements* (and related productivity gains) and *post-editing rates*, measured by the Human-targeted Translation Edit Rate (HTER, Snover et al. 2006). Finally, we assess the subjective perception of PE gains by examining the post-task questionnaires. We reiterate that all scores in this section are computed on the same set of source sentences for all languages, resulting in a faithful cross-lingual comparison of post-editing effort thanks to DivEMT's controlled setup.

4.1 Temporal Effort and Productivity Gains

We start by comparing *task time* (seconds per processed source word) across languages and modalities. For this purpose, edit times are computed for every document in every language without considering the presence of multiple translators for every language. As shown in Figure 2, translation time varies considerably across languages even when no MT system is involved (HT), suggesting an intrinsic variability in translation complexity for different subjects and language pairs. Indeed, for the HT modality, the time required for the 'slowest' target languages (Italian, Ukrainian) is roughly double the 'fastest' one (Turkish). This pattern cannot be easily explained and contrasts with factors commonly tied to MT complexity, such as source-target morphological richness and language relatedness (Birch et al., 2008; Belinkov et al., 2017). On the other hand, we find the relation $PE_1 > PE_2 > HT$ (PE_1 fastest, PE_2 medium speed, HT slowest) to hold for all the evaluated languages.

	PROD ↑			ΔHT ↑	
	HT	PE_1	PE_2	PE_1	PE_2
ARA	13.1	21.7	16.3	+84%	+10%
NLD	13.6	28.7	21.7	+119%	+61%
ITA	8.8	18.6	15.6	+96%	+95%
TUR	17.9	25.5	21.0	+34%	+12%
UKR	8.0	12.3	9.8	+71%	+14%
VIE	10.2	13.0	11.1	+32%	+23%

Table 4: Median productivity (PROD, # processed source words per minute) and median % post-editing speedup (ΔHT) for all analyzed languages and modalities. Arrows denote the direction of improvement.

For a measure of productivity gains that is easier to interpret and more in line with translation industry practices, we turn to *productivity* expressed in source words processed per minute and compute the *speed-up* induced by the two post-editing modalities over translating from scratch (ΔHT). Table 4 presents our results. **Across systems**, we find that *large* differences among automatic MT quality metrics indeed reflect on post-editing effort, suggesting a nuanced picture that is complementary to the findings of Zouhar et al. (2021). While post-editing time gains were observed to quickly saturate for slight changes in high-quality MT, we find that moving from medium-quality to high-quality MT yields meaningful productivity improvements across most evaluated languages. **Across languages**, too, the magnitude of productivity gains ranges widely, from doubling in some languages (Dutch PE_1, Italian PE_1 and PE_2) to only about 10% (Arabic, Turkish and Ukrainian PE_2). When only considering the better performing system (PE_1), post-editing remains clearly beneficial in all languages despite the high variability in ΔHT scores. Results are more nuanced for the open-source system (PE_2), with three out of six languages displaying only marginal gains (<15% in Arabic, Turkish and Ukrainian). Despite its overall lower performance, mBART-50 (PE_2) is the only system enabling a fair comparison across languages (from the point of view of training data size and architecture, see Section 3.4). Interestingly, if we focus on the gains induced by this system, factors like language relatedness and morphological complexity become relevant. Specifically, Italian (+95%), Dutch (+61%) and Ukrainian (+14%) are genetically and syntactically related to English, but Ukrainian has a richer morphology (see Table 1). On the other hand, Vietnamese (+23%), Turkish

(+12%) and Arabic (+10%) all belong to different families. However, Vietnamese is isolating (little to no morphology), while Turkish and Arabic have very rich morphological systems (resp. agglutinative and introflexive, the latter of which is especially problematic for subword segmentation, Amrhein and Sennrich 2021). Other differences are however harder to explain. For instance, Dutch is closely related to English and has a simpler morphology than Italian, but its productivity gain with mBART-50 is lower (61% vs 95%). This finding is accompanied by an important gap in BLEU and COMET scores achieved by mBART-50 on the two languages (22.6 vs 24.4 BLEU and 0.532 vs 0.648 COMET for Dutch vs Italian, resp.) which cannot be explained by training data size.

In summary, our findings confirm the overall positive impact of NMT post-editing on translation productivity observed in previous PE studies. However, we note how **the magnitude of this impact is highly variable across systems and languages**, with inter-subject variability also playing an important role, in line with previous studies (Koponen et al., 2020) (see Section 6 for more details). The small size of our language sample does not allow us to draw direct causal links between specific typological properties and post-editing efficiency. That said, we believe these results have important implications on the claimed 'universality' of current state-of-the-art MT and NLP systems, mostly based on the Transformer architecture (Vaswani et al., 2017) and BPE-style subword segmentation techniques (Sennrich et al., 2016).

4.1.1 Modeling Temporal Effort

Given the high variability among translators, segments and translation modalities, we assess the validity of our observations via statistical analysis of temporal effort using a linear mixed-effects regression model (LMER, Lindstrom and Bates 1988), following Green et al. (2013) and Toral et al. (2018). We fit our model on $n = 7434$ instances, corresponding to 413 sentences translated by 18 translators[8], using translation time as the dependent variable. Our fixed predictors include translation modality, target language, their interaction and length of source segment in characters.[9] Our random effects structure includes random intercepts for different segments (nested with documents) and translators, as well as a random slope for modality over individual segments.[10] Table 5 presents the set of predictors included in the final model, an estimate of their impact on edit times and their significance. We find both PE modalities to significantly reduce translation times ($p < 0.001$), with PE_1 being significantly faster than PE_2 ($p < 0.001$) across all languages. Taking the language for which HT is slowest (Ukrainian) as the reference level, the reduction in time brought by Google is significantly more pronounced for Italian, Dutch ($p < 0.001$), and Turkish ($p < 0.05$). For mBART-50, however, we only observe significantly more pronounced increases in productivity for Italian and Dutch ($p < 0.001$) compared to the reference. We find these results to corroborate the observations of the previous section.

Predictor	Estim.	p-value	Sig.
(intercept)	4.92	1.12e-11	***
source length	0.38	< 2e-16	***
lang_ara	-0.49	0.1209	
lang_ita	-0.14	0.6407	
lang_nld	-0.58	0.0733	x
lang_tur	-0.82	0.0162	*
lang_vie	-0.24	0.4254	
task_pe1	-0.49	< 2e-16	***
task_pe2	-0.22	1.77e-07	***
lang_ara:task_pe1	-0.11	0.0505	x
lang_ita:task_pe1	-0.40	8.97e-12	***
lang_nld:task_pe1	-0.41	5.74e-12	***
lang_tur:task_pe1	-0.14	0.0194	*
lang_vie:task_pe1	0.13	0.0290	*
lang_ara:task_pe2	0.05	0.3535	
lang_ita:task_pe2	-0.39	3.30e-11	***
lang_nld:task_pe2	-0.29	4.46e-07	***
lang_tur:task_pe2	0.03	0.5811	
lang_vie:task_pe2	0.04	0.5289	

Table 5: LMER modeling results using translation time as the dependent variable. The reference levels for predictors lang and task are Ukrainian and Translation from scratch (HT), respectively. Estimate impact on edit time for every predictor is provided in log seconds. Significance: *** = < 0.001, * = < 0.05, x = < 0.1

4.2 Post-Editing Rate

We proceed to study the post-editing patterns using the widely-adopted Human-targeted Translation Edit Rate (HTER, Snover et al. 2006), computed as the length-normalized sum of word-level substitutions, insertions, deletions and shift operations

[8] Outliers were removed beforehand, see Appendix D.

[9] The document processing order was originally included to identify possible longitudinal effects but was removed due to a lack of significant improvements.

[10] Additional modeling details available in Appendix G.

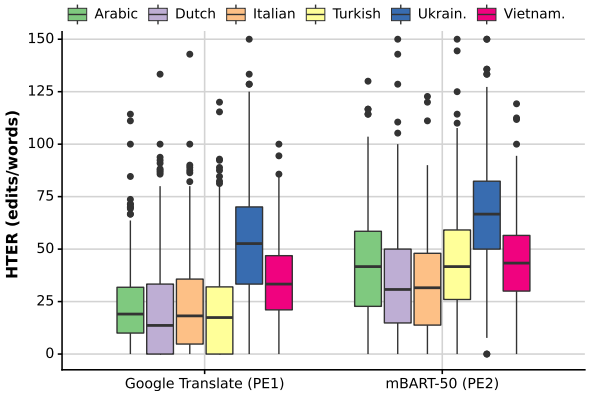

Figure 3: Human-targeted Translation Edit Rate (HTER) for Google Translate and mBART-50 post-editing across available languages.

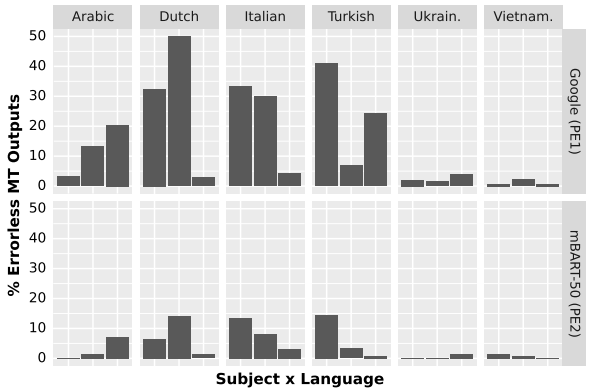

Figure 4: Distribution of error-less machine translation sentence outputs (no edits performed during post-editing) for each translator and every language.

performed during post-editing.[11]

As shown in Figure 3, PE_1 required less editing than PE_2 for all languages, and a high variability is observed across the two systems and all languages. Since translators were not informed about the presence of two MT systems, we exclude the possibility that these results reflect an over-reliance or distrust towards a specific MT system. For Google Translate, Ukrainian shows the heaviest edit rate, followed by Vietnamese, whereas Arabic, Dutch, Italian and Turkish all show relatively low amounts of edits. Focusing again on mBART-50 for a fairer cross-lingual comparison, Ukrainian is by far the most heavily edited language, followed by a medium-tier group composed of Vietnamese, Arabic and Turkish, and finally by Dutch and Italian as low-edit languages. Results show that several of our observations on the linguistic relatedness and type of morphology also apply to edit rates, with

[11] See Appendix E for extra results with a character-level variant of HTER.

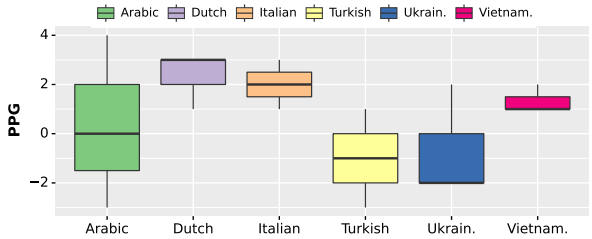

Figure 5: Perceived productivity gains (PPG) between the HT and PE translation modalities, assessed for all subjects after task completion.

languages less related to English or having richer morphology requiring more post-edits on average.

Figure 4 visualizes the large gap in edit rates across languages and subjects by presenting the amount of "errorless" MT sentences that were accepted directly, i.e. without any post-editing. We note again how the NMT system heavily influences the rate of occurrence of such sentences but nonetheless shows how Dutch and Italian generally present more errorless sentences than Ukrainian and Vietnamese. In particular, for Google Translate outputs, the average rate of error-less sentences is roughly 25% for the former target languages, while for the latter, it accounts only for the 3% of total translations. Surprisingly, the English-Turkish pair also fares well, despite the low source-target relatedness.

Finally, we note that post-editing effort appears to correlate poorly with the automatic MT quality metrics reported in Table 2 (e.g. see high scores of Vietnamese and low scores of Dutch PE_1), highlighting a difficulty in predicting the benefits of MT post-editing over HT for new language pairs.

4.3 Perception of Productivity Gain

We conclude our analysis by examining the post-task questionnaires, in which participants expressed their perception of MT quality and translation speed across HT and PE modalities (HT_s, PE_s)[12] using a 1-7 Likert scale (1 slowest, 7 fastest). We use these to compute the Perceived Productivity Gain (PPG) as $PPG = PE_s - HT_s$ and visualize it in Figure 5. We observe that Italian and Dutch, the only target languages with marked productivity gains (ΔHT) regardless of the PE system in Table 4, are also the only ones having consistently high (≥ 2) PPG scores across all subjects. Moreover, we remark how PPG for target languages

[12] We reemphasize that subjects were unaware of the presence of two distinct MT systems.

with a large gap in ΔHT scores between high-PE_1 and low-PE_2 (Arabic, Ukrainian) are hardly distinguishable from those of languages in which ΔHT is low for both PE systems (Turkish, Vietnamese). Notably, 4 out of 18 subjects attribute negative PPGs to the PE modality, even though productivity gains were reported across all subjects and languages. These results suggest that worst-case usage scenarios may play an important role in driving PPG, i.e. that *subjects' perception of quality is largely shaped by particularly challenging or unsatisfying interactions with the NMT system, rather than the average case*. Finally, from the post-task questionnaire, PPG scores exhibit a strong positive correlation with the perception of MT adequacy (ρ=0.66), fluency (ρ=0.46) and overall quality (ρ=0.69), and more generally with a higher enjoyability of PE (ρ=0.60), while being inversely correlated with the perception of problematic mistranslations (ρ=−0.60).

5 Conclusions

In this work we introduced DivEMT, the outcome of a post-editing study spanning two state-of-the-art NMT systems, 18 professional translators and six typologically diverse target languages under a unified setup.

We leveraged DivEMT's behavioral data to perform a controlled cross-language analysis of NMT post-editing effort along its temporal and editing effort dimensions. The analysis reveals that NMT drives significant improvements in productivity across all the evaluated languages, but the magnitude of these improvements depends heavily on the language and the underlying NMT system. In this setting, productivity measurements across modalities were found to be generally consistent with the recorded editing patterns. Our results indicate that translators working on language pairs with significant post-editing productivity gains, on average, perform fewer edits and accept more machine-generated translations without any editing. We also observed a disconnect between post-editing productivity gains and MT quality metrics collected for the same NMT systems. Finally, low source-language relatedness and target morphological complexity seem to hinder productivity when NMT is adopted, even in settings where system architecture and training data are controlled for.

In our qualitative analysis, translators' perception of post-editing usefulness was found to be strongly shaped by problematic mistranslations. Languages showing large productivity gains for both NMT systems were the only ones associated with a positive perception of PE-mediated gains, as opposed to mixed or negative opinions for other translation directions.

In future work, a more fine-grained analysis of the types of edits conducted by the translators, and their differences across languages, could shed more light on our current findings.

6 Limitations

The subjective component introduced by the presence of multiple translators is an important confounding factor in our setup, especially due to the relatively small number of subjects for each language. In our study, we tried to balance a thorough control of other noise components with a faithful reproduction of a realistic translation scenario. However, we realize that the combination of limited document context provided by FLORES-101, the variety of topics covered in the texts and the experimental nature of the PET platform constitutes an atypical setting that may have impacted the translators' natural productivity. Moreover, variability in the content of mBART-50 fine-tuning data, despite the comparable sizes, may have played a role in the variability observed for automatic MT evaluation and PE gains across languages.

7 Broader Impact and Ethical Considerations

This line of research aims at providing a more precise and faceted understanding of translation and editing effort across multiple languages, and as such is worth pursuing to ensure a fairer compensation to translators if compared to one-size-fits-all approaches based on automatic quality metrics. Furthermore, the understanding of the application of MT to translators' work in less researched languages and the diversity of measures obtained can give a clearer picture of MT usability, in its broader sense, than automatic metrics. It is relevant to test NMT models in controlled translation environments. In our experiment, Language Service Providers were paid their requested rate. All words were paid as new words, as the MT usability was unknown prior to the experiment. They were also given thorough instructions and ample time to complete the assignment, accommodating for the COVID-19 pandemic that affected some of the

participants. Translators were informed that they could opt-out at any time and have their information deleted.

Acknowledgements

Data collection was fully funded by the Dutch Research Council (NWO) under project number 639.021.646. GS is supported by the NWO project InDeep (NWA.1292.19.399). AGA was supported by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 890697. We thank the Center for Information Technology of the University of Groningen for providing access to the Peregrine HPC cluster.

References

Arafat Ahsan, Vandan Mujadia, and Dipti Misra Sharma. 2021. Assessing post-editing effort in the English-Hindi direction. In *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*, National Institute of Technology Silchar , Silchar, India. NLP Association of India (NLPAI).

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online. Association for Computational Linguistics.

Chantal Amrhein and Rico Sennrich. 2021. How suitable are subword segmentation strategies for translating non-concatenative morphology? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 689–705, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wilker Aziz, Sheila Castilho, and Lucia Specia. 2012. PET: a tool for post-editing and assessing machine translation. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3982–3987, Istanbul, Turkey. European Language Resources Association (ELRA).

Loic Barrault, Ondrej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussa, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Tom Kocmi, Andre Martins, Makoto Morishita, and Christof Monz, editors. 2021. *Proceedings of the Sixth Conference on Machine Translation*. Association for Computational Linguistics, Online.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017. What do neural machine translation models learn about morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 861–872, Vancouver, Canada. Association for Computational Linguistics.

Luisa Bentivogli, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. 2016. Neural versus phrase-based machine translation quality: a case study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 257–267, Austin, Texas. Association for Computational Linguistics.

Alexandra Birch, Miles Osborne, and Philipp Koehn. 2008. Predicting success in machine translation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 745–754, Honolulu, Hawaii. Association for Computational Linguistics.

Arianna Bisazza, Ahmet Üstün, and Stephan Sportel. 2021. On the difficulty of translating free-order case-marking languages. *Transactions of the Association for Computational Linguistics*, 9:1233–1248.

Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 1–44, Sofia, Bulgaria. Association for Computational Linguistics.

Emanuele Bugliarello, Sabrina J. Mielke, Antonios Anastasopoulos, Ryan Cotterell, and Naoaki Okazaki. 2020. It's easier to translate out of English than into it: Measuring neural translation difficulty by cross-mutual information. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1640–1649, Online. Association for Computational Linguistics.

Sheila Castilho, Joss Moorkens, Federico Gaspari, Iacer Calixto, John Tinsley, and Andy Way. 2017. Is Neural Machine Translation the New State of the Art? *The Prague Bulletin of Mathematical Linguistics*, 108(1):109–120. Publisher: Sciendo Section: The Prague Bulletin of Mathematical Linguistics.

Mauro Cettolo, Marcello Federico, Luisa Bentivogli, Jan Niehues, Sebastian Stüker, Katsuhito Sudoh,

Koichiro Yoshino, and Christian Federmann. 2017. Overview of the IWSLT 2017 evaluation campaign. In *Proceedings of the 14th International Conference on Spoken Language Translation*, pages 2–14, Tokyo, Japan. International Workshop on Spoken Language Translation.

Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, Rolando Cattoni, and Marcello Federico. 2016. The IWSLT 2016 evaluation campaign. In *Proceedings of the 13th International Conference on Spoken Language Translation*, Seattle, Washington D.C. International Workshop on Spoken Language Translation.

Ryan Cotterell, Sabrina J. Mielke, Jason Eisner, and Brian Roark. 2018. Are all languages equally hard to language-model? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 536–541, New Orleans, Louisiana. Association for Computational Linguistics.

Joke Daems, Sonia Vandepitte, Robert Hartsuiker, and Lieve Macken. 2017. Translation Methods and Experience: A Comparative Analysis of Human Translation and Post-editing with Students and Professional Translators. *Meta : journal des traducteurs / Meta: Translators' Journal*, 62(2):245–270. Publisher: Les Presses de l'Université de Montréal.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Çelebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Marina Fomicheva, Shuo Sun, Erick Rocha Fonseca, F. Blain, Vishrav Chaudhary, Francisco Guzmán, Nina Lopatina, Lucia Specia, and André F. T. Martins. 2020. MLQE-PE: A multilingual quality estimation and post-editing dataset. *ArXiv*, abs/2010.04480.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Daniela Gerz, Ivan Vulić, Edoardo Maria Ponti, Roi Reichart, and Anna Korhonen. 2018. On the relation between linguistic typology and (limitations of) multilingual language modeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 316–327, Brussels, Belgium. Association for Computational Linguistics.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 Evaluation Benchmark for Low-Resource and Multilingual Machine Translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Spence Green, Jeffrey Heer, and Christopher D. Manning. 2013. The efficacy of human post-editing for language translation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 439–448, New York, NY, USA. Association for Computing Machinery.

Ana Guerberof. 2009. Productivity and quality in MT post-editing. In *Beyond Translation Memories: New Tools for Translators Workshop*, Ottawa, Canada.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Philipp Koehn and Christof Monz, editors. 2006. *Proceedings on the Workshop on Statistical Machine Translation*. Association for Computational Linguistics, New York City.

Maarit Koponen, Wilker Aziz, Luciana Ramos, and Lucia Specia. 2012. Post-editing time as a measure of cognitive effort. In *Workshop on Post-Editing Technology and Practice*.

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, Lisboa, Portugal. European Association for Machine Translation.

Hans P. Krings. 2001. *Repairing Texts: Empirical Investigations of Machine Translation Post-editing Processes*. Kent State University Press. Google-Books-ID: vsdPsIXCiWAC.

Isabel Lacruz, Michael Denkowski, and Alon Lavie. 2014. Cognitive demand and cognitive effort in post-editing. In *Proceedings of the 11th Conference of the Association for Machine Translation in the Americas*, pages 73–84, Vancouver, Canada. Association for Machine Translation in the Americas.

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. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 267–272, Dublin, Ireland. European Association for Machine Translation.

Samuel Läubli, Mark Fishel, Gary Massey, Maureen Ehrensberger-Dow, and Martin Volk. 2013. Assessing post-editing efficiency in a realistic translation environment. In *Proceedings of the 2nd Workshop on Post-editing Technology and Practice*, Nice, France.

Daniel Licht, Cynthia Gao, Janice Lam, Francisco Guzman, Mona Diab, and Philipp Koehn. 2022. Consistent human evaluation of machine translation across language pairs. *ArXiv*, abs/2205.08533.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, Florence, Italy. Association for Computational Linguistics.

Mary J. Lindstrom and Douglas M. Bates. 1988. Newton—raphson and em algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association*, 83(404):1014–1022.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Zihan Liu, Genta Indra Winata, and Pascale Fung. 2021. Continual mixed-language pre-training for extremely low-resource neural machine translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2706–2718, Online. Association for Computational Linguistics.

Arl Lommel and Hélène Pielmeier. 2021. Machine Translation Use at LSPs: Insights on How Language Service Providers' Use of MT Is Evolving. Technical report, Common Sense Advisory Research.

Thomas Mayer and Michael Cysouw. 2014. Creating a massively parallel Bible corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3158–3163, Reykjavik, Iceland. European Language Resources Association (ELRA).

Sabrina J. Mielke, Ryan Cotterell, Kyle Gorman, Brian Roark, and Jason Eisner. 2019. What kind of language is hard to language-model? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4975–4989, Florence, Italy. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Carla Parra Escartín and Manuel Arcedillo. 2015. Machine translation evaluation made fuzzier: a study on post-editing productivity and evaluation metrics in commercial settings. In *Proceedings of Machine Translation Summit XV: Papers*, Miami, USA.

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).

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, Lisbon, Portugal. Association for Computational Linguistics.

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, Brussels, Belgium. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Maria Stasimioti and Vilelmini Sosoni. 2020. Translation vs post-editing of NMT output: Insights from the English-Greek language pair. In *Proceedings of 1st Workshop on Post-Editing in Modern-Day Translation*, pages 109–124, Virtual. Association for Machine Translation in the Americas.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2021. Multilingual translation from denoising pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466, Online. Association for Computational Linguistics.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Antonio Toral, Martijn Wieling, and Andy Way. 2018. Post-editing Effort of a Novel With Statistical and Neural Machine Translation. *Frontiers in Digital Humanities*, 5:1–11. Publisher: Frontiers.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Weiyue Wang, Jan-Thorsten Peter, Hendrik Rosendahl, and Hermann Ney. 2016. CharacTer: Translation edit rate on character level. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 505–510, Berlin, Germany. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Aris Xanthos, Sabine Laaha, Steven Gillis, Ursula Stephany, Ayhan Aksu-Koç, Anastasia Christofidou, Natalia Gagarina, Gordana Hrzica, F. Nihan Ketrez, Marianne Kilani-Schoch, Katharina Korecky-Kröll, Melita Kovačević, Klaus Laalo, Marijan Palmović, Barbara Pfeiler, Maria D. Voeikova, and Wolfgang U. Dressler. 2011. On the role of morphological richness in the early development of noun and verb inflection. *First Language*, 31(4):461–479.

Vilém Zouhar, Martin Popel, Ondřej Bojar, and Aleš Tamchyna. 2021. Neural machine translation quality and post-editing performance. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10204–10214, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Çağrı Çöltekin and Taraka Rama. 2022. What do complexity measures measure? correlating and validating corpus-based measures of morphological complexity. *ArXiv*, abs/2204.05056.

A Modality scheduling

Table 6 shows an example of the adopted modality scheduling. The modality of document docM_i for translator T_j in the main task is picked randomly among the two modalities that were not seen by the same translator for docM_{i-1}, enforcing consecutive documents given to the same translator to be assigned different modalities to avoid periodicity in repetition and enable the same-language comparisons of Section 4. Importantly, although all three modes were collected for every document, we did not enforce mode consistency across the same translator identifier across languages (i.e. T_1 for Italian does not have the same sequence of modalities of translator T_1 in Arabic, for example). For this reason, individual subjects are not directly comparable across languages. This is relevant since, e.g. T_3 for Dutch and Italian did not operate on the same set of sentences on the same modalities, and thus their comparable editing behavior in Figure 4 should be attributed to personal preference rather than an identical assignment of modalities on the same sentences. Despite modality scheduling, we have no guarantees that translators consistently follow the order of documents presented in PET, and thus possibly operate on documents assigned to the same modality consecutively. However, this possibility reduces to random guessing due to a lack of any identifying information related to the modality until the document is entered for editing. The sequence of modalities for the warmup task is fixed and is: HT, PE_2, PE_1, HT, PE_2.

B Subject Information

During the setup of our experiment, one translator refused to carry out the main task after the warmup phase, and another was substituted by our choice. Both translators were working in the English-Italian direction and were found to make heavy usage of copy-pasting during the warmup stage, suggesting an incorrect utilization of the platform in light of our guidelines. Both translators, which we identified as T_2 and T_3 for Italian, were replaced by T_5 and T_4 respectively. Table 7 reflects the final translation selection for all languages, with the information collected by means of the pre-task questionnaire.

C Translation Guidelines

An extract of the translation guidelines provided to the translators follows. The full guidelines are provided in the additional materials.

		T_1	T_2	T_3
warm-up	docW_1	HT	HT	HT
	docW_2	PE_1	PE_1	PE_1
	...			
	docW_N	PE_2	PE_2	PE_2
main	docM_1	HT	PE_1	PE_2
	docM_2	PE_2	HT	PE_1
	docM_3	HT	PE_2	$\text{PE}_!$
	...			
	docM_N	PE_2	PE_1	HT

Table 6: Modality scheduling overview. For each language, each subject (T_i) works with a pseudo-random sequence of modalities (HT, PE_1, PE_2). For the warm-up task ($N = 5$), all translators are provided with the same documents in the same modalities. For the main task ($N = 107$), each translator is assigned a modality at random. Each document is translated once for every modality. The same procedure is repeated independently for all the languages.

Fill in the pre-task questionnaire before starting the project. In this experiment, your goal is to complete the translation of multiple files in one of two possible translation settings. Please, complete the tasks on your own, even if you know another translator that might be working on this project. The translation setting alternates between texts, with each text requiring a single translation in the assigned setting. The two translation settings are:

1. Translation from scratch. Only the source sentence is provided, you are to write the translation from scratch.

2. Post-editing. The source sentence is provided alongside a translation produced by an MT system. You are to post-edit this MT output. Post-edit the text so you are satisfied with the final translation (the required quality is publishable quality). If the MT output is too time-consuming to fix, you can delete it and start from scratch. However, please do not systematically delete the provided MT output to give your own translation.

Important: All editing MUST happen in the provided PET interface: that is, working in other editors and copy-pasting the text back to PET is NOT ALLOWED, because it invalidates the experiment. This is easy to spot in the log data, so please avoid doing this. Complete the translation of all files sequentially, i.e. in the order presented in the tool. DO NOT SKIP files at your own convenience. Make sure that ALL files are translated when you deliver the tasks.

The aim is to produce publishable professional quality translations for both translation settings. Thus, please translate to your best abilities. You can return to the files and self-review as many times as you think it is necessary. Important: The time invested to translate is recorded while the active unit (sentence) is in editing mode (yellow background). Therefore:

- Only start to translate when you are in editing mode (yellow background). In other words, do not start thinking how you will translate a sentence when the active unit is not yet in editing mode (green or red background).

- Do not leave a unit in editing mode (yellow background) while you do something else. If you need to do something unrelated in the middle of a translation then go out

		Gender	Age	Degree	Position	En Level	YoE	YoE w/ PE	% PE
Arabic	T_1	M	35-44	BA	Freelancer	C2	> 15	2-5	20%-40%
	T_2	M	25-34	BA	Employed	C2	5-10	2-5	60%-80%
	T_3	M	25-34	MA	Freelancer	C1	5-10	< 2	20%-40%
Dutch	T_1	M	25-34	MA	Freelancer	C2	5-10	5-10	60%-80%
	T_2	F	35-44	MA	Freelancer	C1	10-15	5-10	40%-60%
	T_3	F	25-34	MA	Freelancer	C2	2-5	2-5	20%-40%
Italian	T_1	F	25-34	MA	Employed	C1	5-10	5-10	20%-40%
	T_5	F	25-34	MA	Freelancer	C1	2-5	2-5	40%-60%
	T_4	F	35-44	BA	Freelancer	C2	10-15	5-10	> 80%
Turkish	T_1	F	25-34	BA	Freelancer	C2	5-10	2-5	< 20%
	T_2	F	25-34	BA	Freelancer	C1	5-10	5-10	< 20%
	T_3	M	25-34	High school	Freelancer	C2	10-15	< 2	< 20%
Ukrainian	T_1	F	35-44	MA	Employed	C1	5-10	5-10	20%-40%
	T_2	M	35-44	MA	Employed	C1	10-15	10-15	20%-40%
	T_3	M	35-44	High school	Employed	B2	2-5	2-5	20%-40%
Vietnamese	T_1	F	25-34	MA	Employed	C2	10-15	5-10	40%-60%
	T_2	F	25-34	BA	Freelancer	C1	5-10	< 2	20%-40%
	T_3	F	25-34	MA	Employed	C1	2-5	< 2	< 20%

Table 7: Subjects information for DivEMT. The last three columns represent respectively the number of years of professional experience as a translator (YoE), the number of years of experience with MT post-editing (YoE w/ PE) and the % of work assignments requiring post-editing in the last 12 months (% PE) for each subject.

Type	WN	WV	WB	# Sent.	# Words
3S	11	13	11	105	2168
4S	14	8	13	140	3214
5S	12	13	12	185	3826
Tot.	37	34	36	450	9626

Table 8: Distribution of the selected DivEMT documents across sizes and Wikipedia categories. A Type value of NS stands for documents composed by N contiguous sentences, WN, WV and WB stand respectively for WikiNews, WikiVoyage and Wikibooks

of editing mode and come back to editing mode when you are ready to resume translating.

- First you will be translating a warmup task, and then the main task. When you are translating each file, you can consult the Source text (ST) by looking up the url in the Excel files that we have sent for reference.

In order to find the correct terminology for the translation you can consult any source in the Internet. Important: However, it is NOT ALLOWED to use any MT engine to find terms or alternatives to translations (such as Google Translate, DeepL, MS Translator or any MT engine available in your language). Using MT engines invalidates the experiment, and will be detected in the log data. Please fill-in the post-task questionnaire ONLY ONCE after completing all the translation tasks (both warmup and main tasks).

D Details on Document Selection and Preprocessing

Document selection Table 8 present the distribution of selected documents from the Flores-101 devtest split based on their domain and the number of sentences that compose them. The first goal in the selection process was to preserve a rough balance between the three categories while including mostly 4 and 5-sentence docs which are faster to edit in PET (no need to frequently close and reopen an editing window). Another objective of the selection was to minimize the chance of translators finding the translated version of the Wikipedia article from which documents were taken and copied from there, despite our guidelines. We thus scrape the articles from Wikipedia and assess the number of available translations. Among the selected documents, only a small subset has translations in other languages (see Figure 6 top, an article can have multiple languages), mainly in Hebrew (14), Chinese (10), Spanish (7) and German (5) respectively. Considering the total number of translations for every article (Figure 6 bottom), we see that roughly 75% of them (79 docs) have no translations. We consider this satisfactory as proof there should not be a large amount of possible copying involved, and we follow up on this evaluation by also ensuring that no repeated copy-paste patterns are present in keylogs after the warmup stage.

Filtering of Outliers For our analysis of Section 4, we only use sentences with an editing time lower than 45 minutes, which was selected heuristically as a reasonably high threshold to allow for extensive searching and thinking. In the following,

Field name	Description
unit_id, flores_id, subject_id, task_type	Identifiers for the item, respective FLORES-101 sentence, translator and translation mode.
src_text	The original source sentence extracted from Wikinews, wikibooks or wikivoyage.
mt_text	MT output sentence before post-editing, present only if task_type is 'pe'.
tgt_text	Final sentence produced by the translator (either from scratch or post-editing mt_text)
aligned_edit	Aligned visual representation of the machine translation and its post-edit with edit operations
edit_time	Total editing time for the translation in seconds.
k_letter, k_digit, k_white, k_symbol, k_nav	Number of keystrokes for various key types (letters, digits, keystrokes, whitespaces, punctuation, navigation keys) during the translation.
k_erease, k_copy, k_paste, k_cut, k_do	Number of keystrokes for erease (backspace, cancel), copy, paste, cut and Enter actions during the translation.
k_total	Total number of all keystroke categories during the translation.
n_pause_geq_N, len_pause_geq_N	Number and length of pauses longer than 300ms and 1000ms during the translation.
num_annotations	Number of times the translator focused the target sentence texbox during the session.
n_insert, n_delete, n_substitute, n_shift, tot_shifted_words, tot_edits, hter	Granular editing metrics and overall HTER computed using the Tercom library.
cer	Character-level HTER score computed between the MT and post-edited outputs.
bleu, chrf	Sentence-level BLEU and ChrF scores between MT and post-edited fields computed using the SacreBLEU library with default parameters.
time_per_char, key_per_char, words_per_hour, words_per_minute	Edit time per source character, expressed in seconds. Proportion of keys per character needed to perform the translation. Amount of source words translated or post-edited per hour/minute
subject_visit_order	Id denoting the order in which the translator accessed documents in the interface.

Table 9: Description of the main fields associated to every DivEMT data entry. An entry correspond to a translation in a specific modality (HT, PE_1 or PE_2) for one of the six target languages

we present the identifiers of the sentences that were filtered out during this process. E.g. 54.1 means the first sentence of document 54, having item_id equal to flores101-main-541 in the dataset. Note that the sentences were outliers only for 2/6 languages and were all different, indicating no systematic issues in the sample: ARA: 54.1, 100.3, VIE: 3.1, 3.2, 24.3, 28.4, 33.1, 33.2, 40.3, 41.2, 50.3, 100.1, 102.1, 106.1, 107.2, 107.4. The 17 sentences were removed for all modalities and languages in the analysis of Section 4 to preserve the validity of our comparison, representing a loss of roughly 4% of the total available data, a tolerable amount for our analysis.

Fields Description Table 9 presents the set of fields that were collected for every entry of the DivEMT dataset. The fields related to keystrokes, times, pauses, annotations and visit order were extracted from the event log of PET .per files, while edits information and other MT quality metrics were computed in a second moment with the help of widely-used libraries.

Additional Notes on PET The PET platform was modified to enable a correct right-to-left language visualization, which was necessary for Arabic.

E Other Measurements

CharacTER Across Systems and Languages While HTER is a standard metric adopted both in academic and industrial settings, we also evaluated its character-level variant CharacTER (Wang et al., 2016) to assess whether it could better account for the editing process of morphologically rich languages. Figure 7 presents the CharacTER results. When comparing this plot to the HTER one (Figure 3), we notice that CharacTER preserves the overall trends, but slightly improves the edit rate for Arabic and Turkish with respect to other languages. Nevertheless, we find HTER to correlate slightly better with productivity scores across all tested languages, both at a sentence and at a document level. For this reason, word-level results are reported in the article's main body.

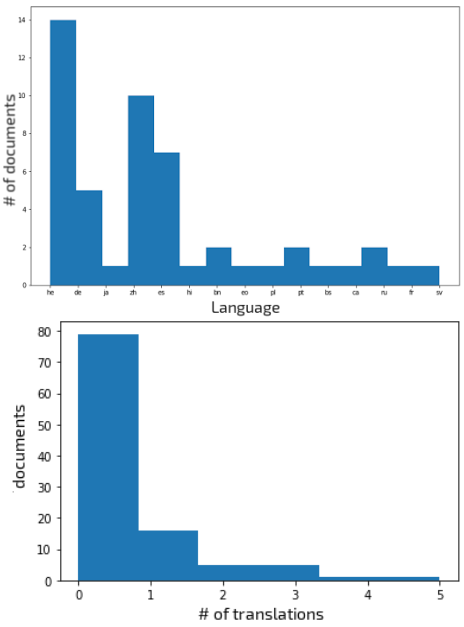

Figure 6: Top: Distribution for the availability of documents selected for DivEMT in languages other than English. Bottom: Quantity of selected documents per number of available translations of Wikipedia.

Automatic Evaluation of NMT Systems The selection of systems used in this study was driven by a broader evaluation procedure covering more models, metrics and target languages. Table 10 presents the overall results of our evaluation. We use HuggingFace's Transformers library (Wolf et al., 2020) for all neural models, using the default decoding settings without further fine-tuning. All metrics were computed using the default settings of SacreBLEU (Post, 2018) and Comet (Rei et al., 2020).

Inter-subject Variability in Translation Times Although the variability across different subjects working on the same language directions is not the main concern of our investigation, we produce Figure 8 (an expanded version of Figure 2) to visualize the inter-subject variability for translation times. We observe that the variability across different translators is more pronounced when translating from scratch and that the overall trend of speed improvements associated with PE is mostly preserved (with few exceptions related to the PE_2 modality).

F Full DivEMT Examples

Tables 11 and 12 present two full examples of DivEMT entries, including all output modalities, intermediate MT outputs, post-edits and edit highlights for all target languages.

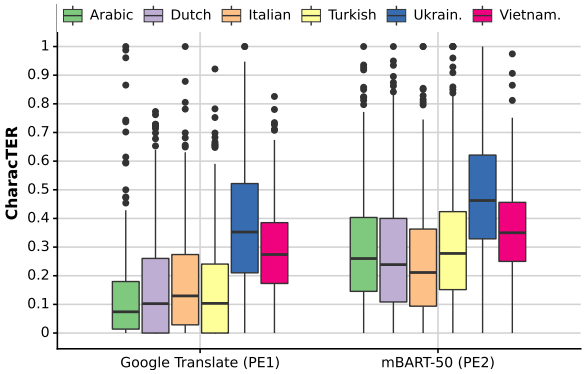

Figure 7: Character-level Human-targeted Translation Edit Rate (CharacTER) for Google Translate and mBART-50 post-editing across available languages.

G Model Description and Feature Significance

Linear Mixed Effects models (LMER) are used for regression analyses involving dependent data, such as longitudinal studies with multiple observations per subject. Given the variables of Table 9, our final model to predict translation time has the following formulation:

```
edit_time ~ src_len_chr + lang_id * task_type
+ (1|subject_id)
+ (1 | document_id/item_id)
+ (0 + task_type | document_id/item_id)
```

We log-transform the dependent variable, edit time in seconds, given its long right tail. The models are built by adding one element at a time, and checking whether such addition leads to a significantly better model with AIC (i.e. if the score gets reduced by at least 2). We fit the models using ML when comparing models that differ in the fixed structure, and REML when they differ in the random structure. We start with an initial model that just includes the two random intercepts (by-translator and by-segment) and proceed by (i) finding significance for nested document/segment random effect; (ii) adding fixed predictors one by one; (iii) adding interactions between fixed predictors; and (iv) adding the random slopes.

From this sequential procedure, we obtain the resulting model. When checking the homoscedasticity and normality of residuals assumptions (Figures 9 and 10), we find the latter is not fulfilled. Consequently, we remove data points for which observations deviate by more than 2.5 standard deviations from the predicted value by the model (2.4% of the data) and refit the best model on this subset,

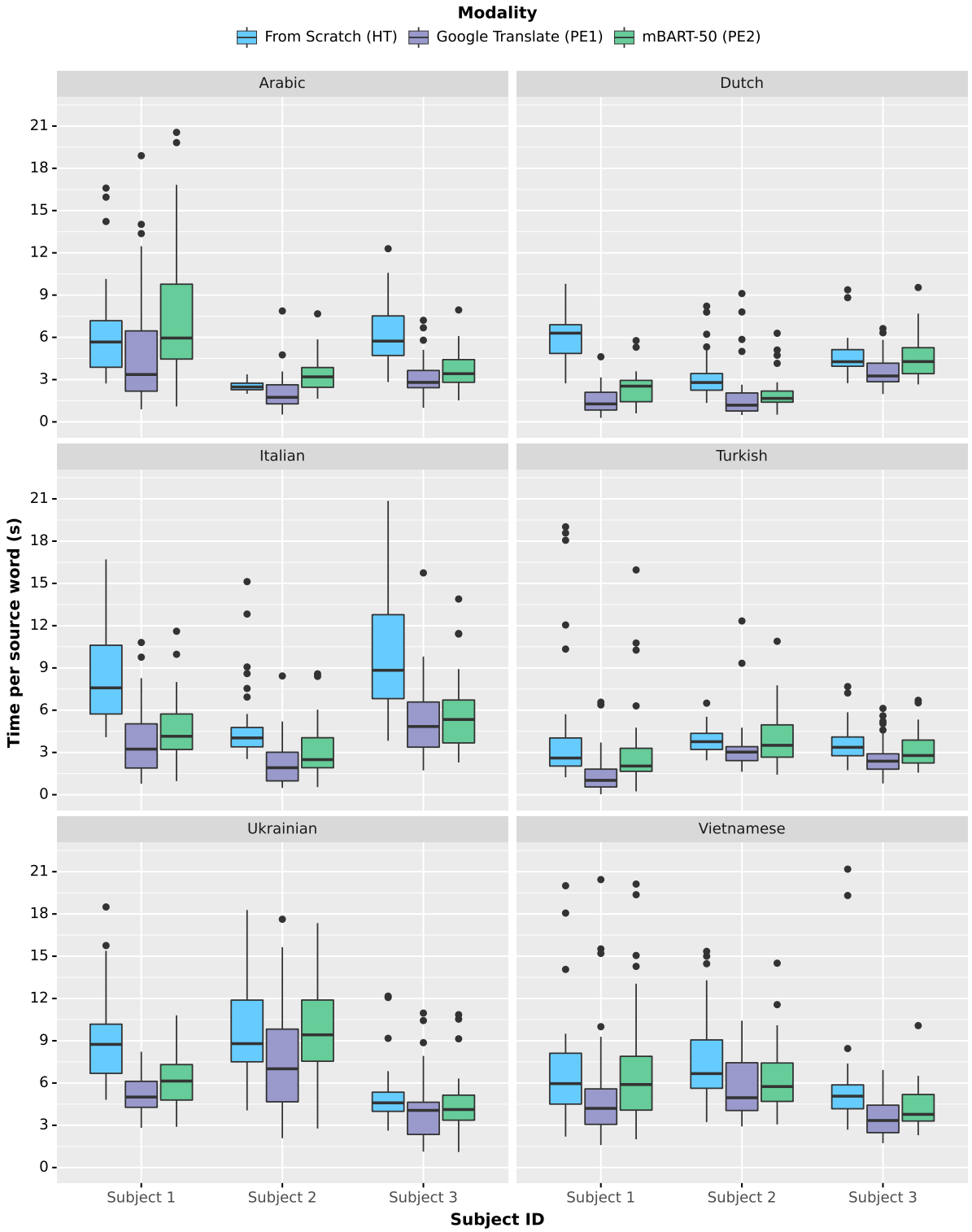

Figure 8: Time per processed source word across languages, subjects and translation modalities, measured in seconds. Each point represents a document containing 3–5 sentences translated by a subject in one of the languages, with higher scores representing slower editing.

	System	BLEU	chrF2	TER	chrF2++	COMET
Arabic	M2M100	19.2	50.9	69.2	47	0.417
	MarianNMT	22.7	54.2	64.7	50.4	0.483
	mBART-50	17	48.5	69.1	44.8	0.452
	GTrans	**34.1**	**65.6**	**52.8**	**61.9**	**0.737**
Dutch	M2M100	21.3	52.9	66.1	49.8	0.405
	MarianNMT	25	56.9	62.5	53.8	0.543
	mBART-50	22.6	53.9	63.7	50.9	0.532
	DeepL	28.7	59.5	59.5	56.6	**0.67**
	GTrans	**29.1**	**60**	**58.5**	**57.1**	0.667
Indonesian	M2M100	35.9	63.1	47.3	60.8	0.614
	MarianNMT	38.5	65.6	46.5	63.3	0.671
	mBART-50	35.9	63.3	47.7	61.1	0.706
	GTrans	**51.5**	**73.6**	**34.5**	**71.9**	**0.894**
Italian	M2M100	23.6	53.9	63.2	51	0.51
	MarianNMT	27.5	57.6	58.9	54.8	0.642
	mBART-50	24.4	54.7	61.2	51.8	0.648
	DeepL	**33**	61	54	58.5	**0.795**
	GTrans	32.8	**61.4**	**53.6**	**58.8**	0.781
Japanese	M2M100	24.5	32.2	123.3	26	0.389
	mBART	27.1	35.4	123	28.3	0.538
	DeepL	**41.3**	**46.8**	108	**37**	**0.75**
	GTrans	38.4	44.7	**101.5**	33.9	0.683
Polish	M2M100	16.1	46.5	74.2	43.1	0.486
	MarianNMT	19.3	49.9	70.5	46.6	0.648
	mBART-50	17.4	48.2	72.4	44.9	0.603
	DeepL	24	54.3	66.4	51.1	**0.832**
	GTrans	**24.4**	**54.6**	**64.6**	**51.4**	0.804
Russian	M2M100	22.5	51.1	65.6	48.1	0.427
	MarianNMT	25.4	53.5	64.3	50.7	0.537
	mBART	24.8	52.6	63.7	49.7	0.541
	DeepL	**35.9**	**61.8**	**53.3**	**59.3**	**0.79**
	GTrans	33	60.5	55.2	57.7	0.731
Turkish	M2M100	20.3	53.9	65.2	50.1	0.686
	MarianNMT	26.3	59.8	58.8	55.8	0.881
	mBART-50	18.8	52.7	67.5	48.7	0.755
	GTrans	**35**	**65.5**	**50.4**	**62.2**	**1**
Ukrainian	M2M100	21.9	51.4	65.8	48.3	0.463
	MarianNMT	20	48.8	69.2	45.7	0.427
	mBART-50	21.9	50.7	67.9	47.7	0.587
	GTrans	**31.1**	**59.8**	**55.9**	**56.8**	**0.758**
Vietnamese	M2M100	33.3	52.3	52.4	52.1	0.43
	MarianNMT	26.7	45.7	60.2	45.6	0.117
	mBART-50	34.7	54	50.7	53.8	0.608
	GTrans	**45.1**	**61.9**	**41.8**	**61.9**	**0.724**

Table 10: Automatic MT quality of all evaluated NMT systems on all tested languages in the English-to-XX setting, using the FLORES-101 full devtest for evaluation. Besides mBART-50 and Google Translate (GTrans), we also evaluate a set of bilingual Transformer-based NMT models trained with MarianNMT (Tiedemann and Thottingal, 2020), the DeepL industrial MT system and the multilingual M2M-100 418M model (Fan et al., 2021). Overall best performance per language is highlighted in **bold**, best open-source system performance per language is underlined.

in order to find out whether any of the effects were due to these outliers. The resulting trends do not change significantly in this final model, in which residuals are normally distributed. As a final sanity check, in Table 13 we measure the effect of subject identity on edit times and find no systematic patterns across languages.

ENGLISH	
	Inland waterways can be a good theme to base a holiday around.
ARABIC	
HT	يمكن أن تكون الممرات المائية الداخلية خياراً جيدًا لتخطيط عطلة حولها.
PE_1	**MT:** يمكن أن تكون الممرات المائية الداخلية موضوعًا جيدًا لإقامة عطلة حولها.
	PE: يمكن أن تكون الممرات المائية الداخلية مظهرًا جيدًا لإقامة عطلة حولها.
PE_2	**MT:** يمكن أن تكون السكك الحديدية الداخلية موضوعًا جيدًا لإقامة عطلة حول.
	PE: قد تكون الممرات المائية الداخلية مكانًا جيدًا لقضاء عطلة حولها.
DUTCH	
HT	Binnenlandse waterwegen kunnen een goed thema zijn voor een vakantie.
PE_1	**MT:** De binnenwateren kunnen een goed thema zijn om een vakantie omheen te baseren .
	PE: Binnenwateren kunnen een goede vakantiebestemming zijn .
PE_2	**MT:** Binnenwaterwegen kunnen een goed thema zijn om een vakantie rond te zetten .
	PE: Binnenwaterwegen kunnen een goed thema zijn om een vakantie rond te organiseren .
ITALIAN	
HT	I corsi d'acqua dell'entroterra possono essere un ottimo punto di partenza da cui organizzare una vacanza.
PE_1	**MT:** Trasporto fluviale può essere un buon tema per basare una vacanza in giro .
	PE: I canali di navigazione interna possono essere un ottimo motivo per cui intraprendere una vacanza .
PE_2	**MT:** I corsi d'acqua interni possono essere un buon tema per fondare una vacanza.
	PE: I corsi d'acqua interni possono essere un buon tema su cui basare una vacanza.
TURKISH	
HT	İç bölgelerdeki su yolları, tatil planı için iyi bir tema olabilir.
PE_1	**MT:** İç su yolları, bir tatili temel almak için iyi bir tema olabilir.
	PE: İç su yolları, bir tatil planı yapmak için iyi bir tema olabilir.
PE_2	**MT:** İç suyolları, tatil için uygun bir tema olabilir.
	PE: İç sular tatil için uygun bir tema olabilir.
UKRAINIAN	
HT	Можна спланувати вихідні, взявши за основу подорож внутрішніми водними шляхами.
PE_1	**MT:** Внутрішні водні шляхи можуть стати гарною темою для відпочинку навколо .
	PE: Внутрішні водні шляхи можуть стати гарною темою для проведення вихідних .
PE_2	**MT:** Водні шляхи можуть бути хорошим об 'єктом для базування відпочинку навколо .
	PE: Місцевість навколо внутрішніх водних шляхів може бути гарним вибором для організації відпочинку.
VIETNAMESE	
HT	Du lịch trên sông có thể là một lựa chọn phù hợp cho kỳ nghỉ.
PE_1	**MT:** Đường thủy nội địa có thể là một chủ đề hay để tạo cơ sở cho một kỳ nghỉ xung quanh .
	PE: Đường thủy nội địa có thể là một ý tưởng hay để lập kế hoạch cho kỳ nghỉ.
PE_2	**MT:** Các tuyến nước nội địa có thể là một chủ đề tốt để xây dựng một kì nghỉ.
	PE: Du lịch bằng đường thủy nội địa là một ý tưởng nghỉ dưỡng không tồi.

Table 11: An example sentence (81.1) from the DivEMT corpus, with the English source and all output modalities for all target languages, including intermediate machine translations (MT) and subsequent post-editings (PE). Colors denote insertions, deletions, substitutions and shifts computed with Tercom (Snover et al., 2006).

ENGLISH	
	The Internet combines elements of both mass and interpersonal communication.
ARABIC	
HT	يجمع الإنترنت بين عناصر وسائل الاتصال العامة والشخصية على حدٍ سواء.
PE_1	**MT:** تجمع الإنترنت بين عناصر الاتصال الجماهيري والشخصي.
	PE: يجمع الإنترنت بين عناصر الاتصال الجماهيري والشخصي.
PE_2	**MT:** إنترنت تجمع عناصر التواصل الجماعي والتواصل الشخصي.
	PE: تجمع شبكة الإنترنت عناصر التواصل الجماعي والتواصل الشخصي.
DUTCH	
HT	Het internet combineert elementen van zowel massa- en intermenselijke communicatie.
PE_1	**MT:** Het internet combineert elementen van zowel massa- als interpersoonlijke communicatie.
	PE: Het internet combineert elementen van zowel massa- als interpersoonlijke communicatie.
PE_2	**MT:** Het internet combineert elementen van massa- en interpersoonlijke communicatie.
	PE: Het internet combineert elementen van massa- en interpersoonlijke communicatie.
ITALIAN	
HT	Internet combina elementi di comunicazione sia di massa sia interpersonale.
PE_1	**MT:** Internet combina elementi di comunicazione di massa e interpersonali .
	PE: Internet combina elementi di comunicazione di massa e interpersonale .
PE_2	**MT:** Internet combina elementi di comunicazione di massa e interpersonale.
	PE: Internet combina elementi di comunicazione di massa e interpersonale.
TURKISH	
HT	İnternet hem kitlesel hem de bireysel iletişim öğelerini birleştiriyor.
PE_1	**MT:** İnternet, hem kitle hem de kişiler arası iletişimin unsurlarını birleştirir.
	PE: İnternet, hem kitleler hem de kişiler arası iletişimin unsurlarını birleştirir.
PE_2	**MT:** İnternet hem kitlesel hem de kişisel iletişim unsurlarını birleştiriyor.
	PE: İnternet hem kitlesel hem de kişisel iletişim unsurlarını birleştiriyor.
UKRAINIAN	
HT	В інтернеті поєднуються елементи групового спілкування та особистого спілкування.
PE_1	**MT:** Інтернет поєднує в собі елементи як масового, так і міжособистісного спілкування.
	PE: Інтернет поєднує в собі елементи як масового, так і міжособистісного спілкування.
PE_2	**MT:** Інтернет об ’єднує як масову , так і міжлюдську комунікацію.
	PE: Інтернет поєднує в собі елементи як групової , так і особистої комунікації.
VIETNAMESE	
HT	Internet là nơi tổng hợp các yếu tố của cả phương tiện truyền thông đại chúng và giao tiếp liên cá nhân.
PE_1	**MT:** Internet kết hợp các yếu tố của cả giao tiếp đại chúng và giao tiếp giữa các cá nhân.
	PE: Internet kết hợp các yếu tố của cả truyền thông đại chúng và giao tiếp giữa các cá nhân.
PE_2	**MT:** Internet kết hợp những yếu tố của sự giao tiếp quần chúng và giao tiếp giữa người với người .
	PE: Internet kết hợp những yếu tố của cả việc giao tiếp đại chúng và giao tiếp cá nhân .

Table 12: An example sentence (29.2) from the DivEMT corpus, with the English source and all output modalities for all target languages, including intermediate machine translations (MT) and subsequent post-editings (PE). Colors denote insertions , deletions , substitutions and shifts computed with Tercom (Snover et al., 2006).

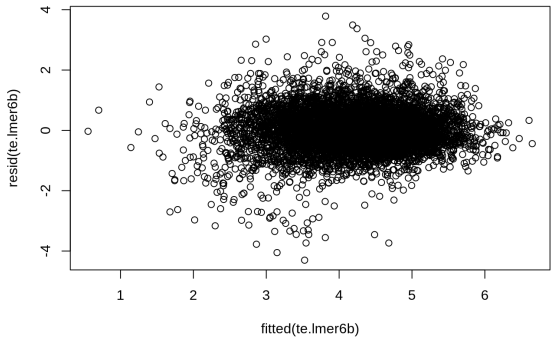

Figure 9: Residuals of the final LMER model, used to verify the heteroscedasticity assumption.

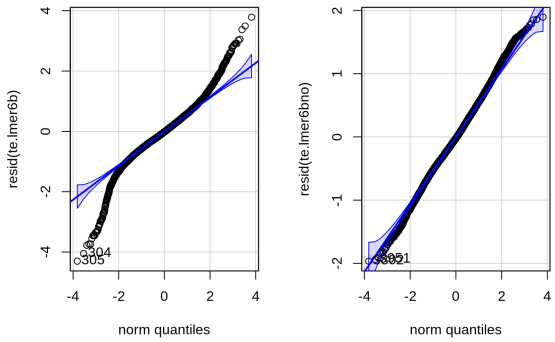

Figure 10: Quantile-quantile plot before and after the removal of outliers when fitting the LMER model, used to verify the normality assumption.

Subject	Coefficient
ara_t1	0.281
ara_t2	-0.384
ara_t3	-0.103
nld_t1	0.001
nld_t2	-0.459
nld_t3	0.458
ita_t1	0.086
ita_t4	0.350
ita_t5	-0.436
tur_t1	-0.381
tur_t2	0.272
tur_t3	0.109
ukr_t1	0.077
ukr_t2	0.314
ukr_t3	-0.391
vie_t1	0.012
vie_t2	0.176
vie_t3	-0.188

Table 13: Coefficients of the random intercept related to the `subject_id` variable, representing the identity of the translator performing the translation.