QE4PE: Word-level Quality Estimation for Human Post-Editing

Gabriele Sarti[1] **Vilém Zouhar**[2] **Grzegorz Chrupała**[3]
Ana Guerberof-Arenas[1] **Malvina Nissim**[1] **Arianna Bisazza**[1]
[1]CLCG, University of Groningen, The Netherlands [2]ETH Zürich, Switzerland
[3]CSAI, Tilburg University, The Netherlands
[1]{g.sarti, a.guerberof.arenas, m.nissim, a.bisazza}@rug.nl
[2]vzouhar@inf.ethz.ch [3]grzegorz@chrupala.me

Abstract

Word-level quality estimation (QE) methods aim to detect erroneous spans in machine translations, which can direct and facilitate human post-editing. While the accuracy of word-level QE systems has been assessed extensively, their usability and downstream influence on the speed, quality, and editing choices of human post-editing remain understudied. In this study, we investigate the impact of word-level QE on machine translation (MT) post-editing in a realistic setting involving 42 professional post-editors across two translation directions. We compare four error-span highlight modalities, including supervised and uncertainty-based word-level QE methods, for identifying potential errors in the outputs of a state-of-the-art neural MT model. Post-editing effort and productivity are estimated from behavioral logs, while quality improvements are assessed by word- and segment-level human annotation. We find that domain, language and editors' speed are critical factors in determining highlights' effectiveness, with modest differences between human-made and automated QE highlights underlining a gap between accuracy and usability in professional workflows.

1 Introduction

Recent years have seen a steady increase in the quality of machine translation (MT) systems and their widespread adoption in professional translation workflows (Kocmi et al., 2024a). Still, human post-editing of MT outputs remains a fundamental step to ensure high-quality translations, particularly for challenging textual domains requiring native fluency and specialized terminology (Liu et al., 2024). Quality estimation (QE) techniques were introduced to reduce post-editing effort by automatically identifying problematic MT outputs without the need for human-written reference translations and were quickly integrated into industry platforms (Tamchyna, 2021). *Segment-level* QE models correlate well with human perception of quality (Freitag et al., 2024) and exceed the performance of reference-based metrics in specific settings (Rei et al., 2021; Amrhein et al., 2022, 2023). On the other hand, *word-level* QE methods for identifying error spans requiring revision have received less attention in the past due to their modest agreement with human annotations, despite their promise for more granular and interpretable quality assessment in line with modern MT practices (Zerva et al., 2024). In particular, while the accuracy of these approaches is regularly assessed in evaluation campaigns, research has rarely focused on assessing the impact of such techniques in realistic post-editing workflows, with notable exceptions suggesting limited benefits (Shenoy et al., 2021; Eo et al., 2022). This hinders current QE evaluation practices: By foregoing experimental evaluation with human editors, it is implicitly assumed that word-level QE will become helpful once sufficient accuracy is reached, without accounting for the additional challenges towards a successful integration of these methods in post-editing workflows.

In this study, which we dub QE4PE (**Q**uality **E**stimation for **P**ost **E**diting), we address this gap by conducting a large-scale study with 42 professional translators for the English→Italian and English→Dutch directions to measure the impact of word-level QE on editing quality, productivity, and usability. We aim for a realistic and reproducible setup, employing the high-quality open-source NLLB 3.3B MT model (NLLB Team et al., 2024) to translate challenging documents from biomedical and social media domains. We then conduct a controlled evaluation of post-editing with error spans in four *highlight modalities*, i.e., using highlights derived

Transactions of the Association for Computational Linguistics, vol. 13, pp. 1410–1435, 2025. https://doi.org/10.1162/TACL.a.46
Action Editor: Jessy Li. Submission batch: 2/2025; Revision batch: 6/2025; Published 10/2025.

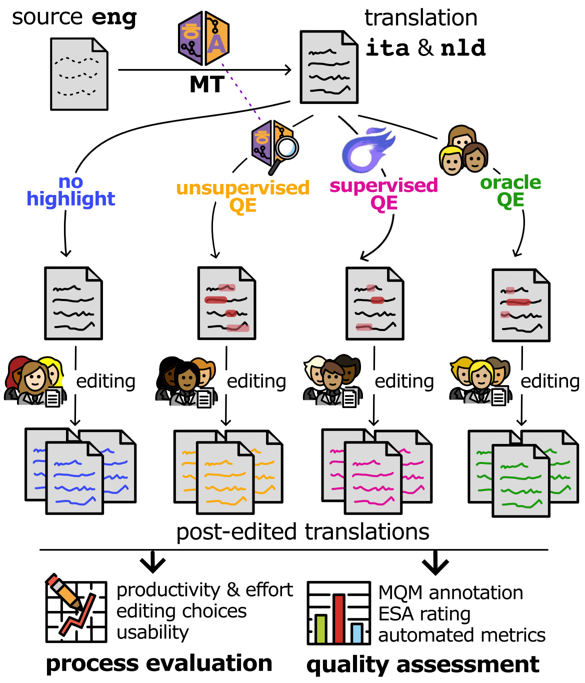


Figure 1: A summary of the QE4PE study. Documents are translated by a neural MT model and reviewed by professional editors across two translation directions and four highlight modalities. Editing effort, productivity, and usability across modalities are estimated from editing logs and questionnaires. Finally, the quality of MT and edited outputs is assessed with MQM/ESA human annotations and automatic metrics.

from four word-level QE methods: a ***supervised*** state-of-the-art QE model trained on human error annotations (XCOMET, Guerreiro et al., 2024), an ***unsupervised*** method leveraging the uncertainty of the MT model during generation, ***oracle*** error spans obtained from the consensus of previous human post-editors, and a ***no highlight*** baseline. The human post-editing is performed using GroTE, a simple online interface we built to support the real-time logging of granular editing data, enabling a quantitative assessment of editing effort and productivity across highlight modalities. We also survey professionals using an online questionnaire to collect qualitative feedback about the usability and quality of the MT model, as well as the interface and error span highlights. Finally, a subset of the original MT outputs and their post-edited variants is annotated following the MQM and ESA protocols (Lommel et al., 2013; Kocmi et al., 2024b) to verify quality improvements after post-editing. See Figure 1 for an overview of the study. Our work represents a step towards an evaluation of translation technologies that is centered on users' experience (Guerberof-Arenas and Moorkens, 2023; Savoldi et al., 2025).

We release all data, code, and the GroTE editing interface to foster future studies on the usability of error span highlighting techniques for other word-level QE methods and translation directions.[1]

2 Related Work

MT Post-Editing Human post-editing of MT outputs is increasingly common in professional translator workflows, as it was shown to increase the productivity of translators while preserving translation quality across multiple domains (Liu et al., 2024). However, many factors were found to influence the variability of post-editing productivity across setups, including MT quality (Zouhar et al., 2021b), interface familiarity (Läubli et al., 2022), individual variability and source-target languages typological similarity (Sarti et al., 2022). Studies evaluating the post-editing process generally focus on *productivity*, i.e., number of processed words/characters per minute, and the temporal, technical and cognitive dimensions of post-editing *effort*, operationalized through behavioral metrics such as editing time, keystrokes and pauses (Krings, 2001; Sarti et al., 2022). We adopt these metrics for the QE4PE study and relate them to different highlight modalities.

Quality Estimation for MT The field of quality estimation was initially concerned with MT model uncertainty (Blatz et al., 2004; Specia et al., 2009), but in time began focusing on predicting translation quality even without using references (Turchi et al., 2013, 2014; Kepler et al., 2019; Thompson and Post, 2020 *inter alia*). Advances in segment- and word-level QE research are regularly assessed in annual WMT campaigns (Fomicheva et al., 2021; Zerva et al., 2022, 2024; Blain et al., 2023), where the best-performing QE systems are usually Transformer-based language models trained on human quality judgments, such as the popular COMET model suite (Rei et al., 2020, 2021, 2022). The widespread adoption of the fine-grained Multidimensional Quality Metrics scale (MQM, Lommel et al., 2013) prompted a paradigm shift in MT evaluation, leading to new QE metrics predicting quality at various granularity levels (Kocmi

[1] Data: `hf.co/gsarti/qe4pe`. Code: `gsarti/qe4pe`.

and Federmann, 2023; Fernandes et al., 2023; Guerreiro et al., 2024). Aside from supervised models, *unsupervised* methods exploiting model uncertainty and its internal mechanisms were proposed as efficient alternatives to identify potential error spans in MT outputs (Fomicheva et al., 2020; Dale et al., 2023; Xu et al., 2023; Himmi et al., 2024, surveyed by Leiter et al. 2024). In this work, we compare the downstream effectiveness of state-of-the-art supervised and unsupervised word-level QE metrics for post-editing settings.

QE for Human Post-Editing Workflows Automatic QE methods are widely used in the translation industry for triaging automatic translations (Tamchyna, 2021). While QE usage has been found helpful to increase the confidence and speed of human assessment (Mehandru et al., 2023; Zouhar et al., 2025), an incautious usage of these techniques can lead to a misplaced over-reliance on model predictions (Zouhar et al., 2021a). Interfaces supporting word-level error highlights were developed for studying MT post-editing (Coppers et al., 2018; Herbig et al., 2020) and code reviewing (Sun et al., 2022; Vasconcelos et al., 2025), with results suggesting that striking the right balance of user-provided information is fundamental to improve the editing experience and prevent cognitive overload. Most similar to our study, Shenoy et al. (2021) investigated the effect of synthetic word-level QE highlights for English→German post-editing on Wikipedia data, concluding that word-level QE accuracy was at the time still insufficient to produce tangible productivity benefits in human editing workflows. In this work, we expand the scope of such evaluation by including two translation directions, two challenging real-world text domains and state-of-the-art MT and QE systems and methods.

3 Experimental Setup

3.1 Structure of the Study

Our study is organized in five stages:

1) Oracle Post-Editing As a preliminary step, segments later used in the main assessment are post-edited by three professionals per direction using their preferred interface without logging. This allows us to obtain post-edits and produce ***oracle*** word-level spans based on the editing consensus of multiple human professionals. Translators involved in this stage are not involved further in the study.

2) Pretask (PRE) The pretask allows the *core translators* (12 per language direction, see Section 3.4) to familiarize themselves with the GROTE interface and text highlights. Before starting, all translators complete a questionnaire to provide demographic and professional information about their profile (Table 10). In the pretask, all translators work in an identical setup, post-editing a small set of documents similar to those of the main task with ***supervised*** highlights. We assign core translators into three groups based on their speed from editing logs (4 translators per group for *faster*, *average*, and *slower* groups in each direction). Individuals from each group are then assigned randomly to each highlight modality to ensure an equal representation of editing speeds, resulting in 1 *faster*, 1 *average*, and 1 *slower* translator for each highlight modality. This procedure is repeated independently for both translation directions.

3) Main Task (MAIN) This task, conducted in the two weeks following the pretask, covers the majority of the collected data and is the main object of study for the analyses of Section 4. In the main task, 24 core translators work on the same texts using the GROTE interface, with three translators per modality in each translation direction, as shown in Figure 1. After the main task, translators complete a questionnaire on the quality and usability of the MT outputs, the interface and, where applicable, word highlights.[2]

4) Post-Task (POST) After MAIN, the 12 core translators per direction are asked to post-edit an additional small set of related documents with GROTE, but this time working all with the ***no highlight*** modality. This step lets us obtain baseline editing patterns for each translator to estimate individual speed and editing differences across highlight modalities without the confounder of interface proficiency accounted for in the PRE stage.

5) Quality Assessment (QA) Finally, a subset consisting of 148 main task segments is randomly selected for manual annotation by six new

[2]We do not disclose the highlight modality to translators to avoid biasing their judgment in the evaluation.

translators per direction (see Section 3.4). For each segment, the original MT output and all its post-edited versions are annotated with MQM error spans, including minor/major error severity and a subset of MQM error categories including e.g., mistranslations, omissions and stylistic errors (Lommel et al., 2013).[3] Moreover, the annotator proposes corrections for each error span, ultimately providing a 0–100 quality score matching the common DA scoring adopted in multiple WMT campaigns. We adopt this scoring system, which closely adheres to the ESA evaluation protocol (Kocmi et al., 2024b), following recent results showing its effectiveness and efficiency for ranking MT system.

In summary, for each translation direction, we collect 3 full sets of oracle post-edits, 12 full sets of edits with behavioral logs for PRE, MAIN, and POST task data, and 13 subsets of main task data (12 post-edits, plus the original MT output) annotated with MQM error spans, corrections and segment-level ESA ratings. Moreover, we also collect 12 pre- and post-task questionnaire responses from *core set* translators to obtain a qualitative view of the editing process.

3.2 Highlight Modalities

We conduct our study on four highlight modalities across two severity levels (*minor* and *major* errors). Using multiple severity levels follows the current MT evaluation practices (Freitag et al., 2021, 2024), and previous results showing that users tend to prefer more granular and informative word-level highlights (Shenoy et al., 2021; Vasconcelos et al., 2025). The highlight modalities we employ are:

No Highlight The text is presented as-is, without any highlighted spans. This setting serves as a baseline to estimate the default post-editing quality and productivity using our interface.

Oracle Following the Oracle Post-editing phase, we produce oracle error spans from the editing consensus of human post-editors. We label text spans that were edited by two out of three translators as *minor*, and those edited by all three translators as *major*, following the intuition that more critical errors are more likely to be identified by several annotators, while minor changes will show more variance across subjects. This modality serves as a best-case scenario, providing an upper bound for future improvements in word-level QE quality.

Supervised In this setting, word-level error spans are obtained using XCOMET-XXL (Guerreiro et al., 2024), which is a multilingual Transformer encoder (Goyal et al., 2021) further trained for joint word- and sentence-level QE prediction. We select XCOMET-XXL in light of its broad adoption, open accessibility and state-of-the-art performance in QE across several translation directions (Zerva et al., 2024). For the severity levels, we use the labels predicted by the model, mapping *critical* labels to the *major* level.

Unsupervised In this modality, we exploit the access to the MT model producing the original translations to obtain *uncertainty-based highlights*. As a preliminary evaluation to select a capable unsupervised word-level QE method, we evaluate two unsupervised QE methods employing token log-probabilities assigned by MT model to predict human post-edits: raw negative log-probabilities (Logprobs), corresponding to the surprisal assigned by the MT model to every generated token, and their variance for 10 steps of Monte Carlo Dropout (MCD Var., Gal and Ghahramani, 2016). We employ surprisal-based metrics following previous work showing their effectiveness in predicting translation errors (Fomicheva and Specia, 2019) and human editing time (Lim et al., 2024). We collect scores for the English→Italian and English→Dutch directions of QE4PE Oracle post-edits and DivEMT (Sarti et al., 2022) to identify the best-performing method, using metric scores extracted from the original models used for translation to predict human post-edits. We use average precision (AP) as a threshold-agnostic performance metric for the tested continuous methods. **Oracle** highlights obtained from the consensus of three annotator in the first stage of the study are used as reference for QE4PE, while a single set of post-edits is available for DivEMT. The XCOMET-XXL model used for **Supervised** highlights, and the average agreement of individual **Oracle** editors with the consensus label are also included for comparison. Table 1[4] show a strong performance for the MCD

[3] See Figure 5 for an overview of setup and guidelines.

[4] Full results in Table 16. Highlights are extended from tokens to words to match the granularity of other modalities.

Method	DivEMT		QE4PE	
	EN-IT	EN-NL	EN-IT	EN-NL
Logprobs	0.18	0.19	0.10	0.09
MCD Var.	**0.41**	**0.42**	**0.23**	**0.31**
XCOMET (**Sup.**)	0.34	0.35	0.16	0.19
Avg. Trans.	–	–	0.53	0.55

Table 1: Average Precision (AP) between metrics and reference error spans.

Task	Domain	# Docs	# Seg.	# Words
PRE	Social	4	23	539
	Biomed.	2	15	348
MAIN	Social	30	160	3375
	Biomed.	21	165	3384
POST	Social	6	34	841
	Biomed.	2	16	257
	Total	64	413	8744

Table 2: Statistics for QE4PE data.

Var. method, even surpassing the accuracy of the supervised XCOMET model across both datasets. Hence, we select MCD Var. for the **Unsupervised** highlight modality, setting value thresholds for minor/major errors to match the respective highlighted word proportions in the **Supervised** modality to ensure a fair comparison.

3.3 Data and MT Model

MT Model On the one hand, the MT model must achieve *high translation quality* in the selected languages to ensure our experimental setup applies to state-of-the-art proprietary systems. Still, the MT model should be *open-source* and have a *manageable size* to ensure reproducible findings and enable the computation of uncertainty for the unsupervised setting. All considered, we use NLLB 3.3B (NLLB Team et al., 2024), a widely used MT model achieving industry-level performances across 200 languages (Moslem et al., 2023).

Data Selection We begin by selecting two translation directions, English→Italian and English→Dutch, according to the availability of professional translators from our industrial partners. We intentionally focus on out-of-English translations as they are generally more challenging for modern MT models (Kocmi et al., 2023). We aim to identify documents that are manageable for professional translators without domain-specific expertise but still prove challenging for our MT model to ensure a sufficient amount of error spans across modalities. Since original references for our selected translation direction were not available, we do not have a direct mean to compare MT quality in the two languages. However, according to our human MQM assessment in Section 4.3 (Table 5), NLLB produces a comparable amount of errors across Dutch and Italian MT, suggesting similar quality.

We begin by translating 3,672 multi-segment English documents from the WMT23 General and Biomedical MT shared tasks (Kocmi et al., 2023; Neves et al., 2023) and MT test suites to Dutch and Italian. Our choice for these specialized domains, as opposed to, e.g., generic news articles, is driven by the real-world needs of the translation industry for domain-specific post-editing support (Eschbach-Dymanus et al., 2024; Li et al., 2025). Moreover, focusing on domains that are considerably more challenging for MT systems than news, as shown by recent WMT campaigns (Neves et al., 2024), ensures a sufficient amount of MT errors to support a sound comparison of word-level QE methods. Then, XCOMET-XXL is used to produce a first set of segment-level QE scores and word-level error spans for all segments. To make the study tractable, we further narrow down the selection of documents according to several heuristics to ensure a realistic editing experience and a balanced occurrence of error spans (details in Appendix A). This procedure yields 351 documents, from which we manually select a subset of 64 documents (413 segments, 8,744 source words per post-editor) across two domains:

- **Social media posts**, including Mastodon posts from the WMT23 General Task (Kocmi et al., 2023) English↔German evaluation and Reddit comments from the Robustness Challenge Set for Machine Translation (RoCS-MT; Bawden and Sagot, 2023), displaying atypical language use, such as slang or acronymization.
- **Biomedical abstracts** extracted from PubMed from the WMT23 Biomedical Translation Task (Neves et al., 2023), including domain-specific terminology.

Table 2 present statistics for the PRE, MAIN, and POST editing stages, and Table 3 shows an example of highlights and edits. While the presence of multiple domains in the same task can render

Source$_{EN}$	So why is it that people jump through extra hoops to install Google Maps?
No High.	Quindi perché le persone devono fare un salto in più per installare Google Maps?
Oracle	Quindi perché le persone devono fare un salto in più per installare Google Maps?
Sup.	Quindi perché le persone devono fare un salto in più per installare Google Maps?
Unsup.	Quindi perché le persone devono fare un salto in più per installare Google Maps?
PE$_{\text{No High.}}$	Quindi perché le persone devono fare un passaggio in più per installare Google Maps?
PE$_{\text{Oracle}}$	Allora, perché le persone fanno un passaggio in più per installare Google Maps?
PE$_{\text{Sup.}}$	Quindi perché le persone fanno passaggi in più per installare Google Maps?
PE$_{\text{Unsup.}}$	Quindi perché le persone fanno i salti mortali per installare Google Maps?

Table 3: EN→IT example from the QE4PE dataset, showing minor/major word highlights and a single post-edit per modality, with modified words highlighted.

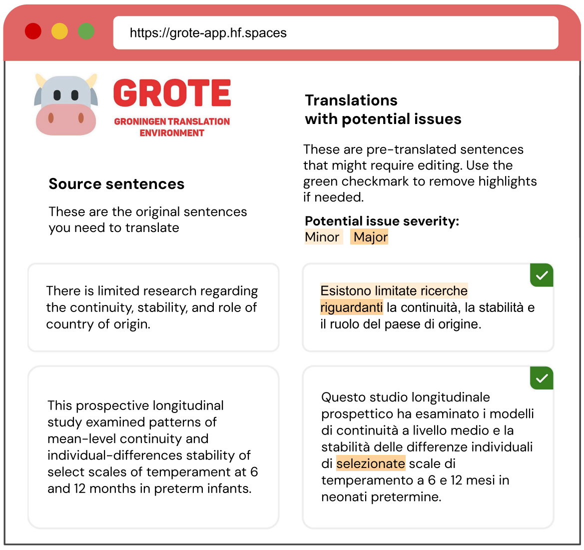


Figure 2: An example of the QE4PE GroTE setup for two segments in an English→Italian document.

our post-editing setup less realistic, we deem it essential to test the cross-domain validity of our findings.

Critical Errors Before producing highlights, we manually introduce 13 critical errors in main task segments to assess post-editing thoroughness. Errors are produced, for example, by negating statements, inverting the polarity of adjectives, inverting numbers, and corrupting acronyms. We replicate the errors in both translation directions to enable direct comparison. Most of these errors were correctly identified across all three highlight modalities (examples in Table 7).

3.4 Participants

For both directions, professional translation companies Translated Srl[5] and Global Textware[6] recruited three translators for the Oracle post-editing stage, the core set of 12 translators working on Pre, Main, and Post tasks, and six more translators for the QA stage, for a total of 21 translators per direction. All translators were freelancers with native proficiency in their target language and self-assessed proficiency of at least C1 in English. Almost all translators had more than two years of professional translation experience and regularly post-edited MT outputs (details in Table 10).

3.5 Editing Interface

We develop a custom interface, which we name **Gro**ningen **T**ranslation **E**nvironment (GroTE, Figure 2), to support editing over texts with word-level highlights. While the MMPE tool used by Shenoy et al. (2021) provides extensive multimodal functionalities (Herbig et al., 2020), we aim for a bare-bones setup to avoid confounders in the evaluation. GroTE is a web interface based on Gradio (Abid et al., 2019) and hosted on the HuggingFace Spaces to enable multi-user data collection online. Upon loading a document, source texts and MT outputs for all segments are presented in two columns following standard industry practices. For modalities with highlights, the interface provides an informative message and supports the removal of all highlights on a segment via a button, with highlights on words disappearing automatically upon editing, as in Shenoy et al. (2021). The interface supports real-time logging of user actions, allowing for the analysis of the editing process. In particular, we log the start/end times for each document, the accessing and exiting of segment textboxes, highlight removals, and keystrokes.

GroTE intentionally lacks standard features such as translation memories, glossaries, and spellchecking to ensure equal familiarity among translators, ultimately controlling for editor proficiency with these tools, as done in previous studies (Shenoy et al., 2021; Sarti et al., 2022). While most translators noted the lack of features in our

[5]https://translated.com.
[6]https://globaltextware.nl/.

usability assessment, the majority also found the interface easy to set up, access, and use (Table 10).

4 Analysis

4.1 Productivity

We obtain segment- and document-level edit times and compute editing *productivity* as the number of processed source characters over the sum of all document-level edit times, measured in characters per minute. To account for potential breaks taken by post-editors during editing, we filter out pauses between logged actions longer than 5 minutes. We note that this procedure does not significantly impact the overall ranking of translators, while ensuring a more robust evaluation of editing time.

Do Highlights Make Post-Editors Faster? Figure 3 shows translators' productivity across stages, with every dot corresponding to the productivity of a single individual. We observe that no highlight modality leads to systematically faster editing across all speed groups and that the ordering of PRE-task speed groups is maintained in the following stages despite the different highlight modalities. These results suggest that individual variability in editing speed is more critical than highlight modality in predicting editing speed. However, faster English→Dutch translators achieve outstanding productivity, i.e., >2 standard deviations above the overall mean (>300 char/min, ➔ in Figure 3) almost exclusively in **No Highlight**, and, **Oracle** modalities, suggesting that lower-quality highlights hinder editing speed.

We validate these observations by fitting a negative binomial mixed-effect model on segment-level editing times (model details in Table 8). Excluding random factors such as translator and segment identity from the model produces a significant drop in explained variance, confirming the inherent variability of editing times ($R^2 = 0.93 \rightarrow 0.41$). Model coefficients show that MT output length and the proportion of highlighted characters are the main factors driving an increase in editing times, possibly reflecting an increase in cognitive effort to process additional information. We find highlights to have a significant impact on increasing the editing speed of English→Italian translators ($p < 0.001$), but a minimal impact for English→Dutch. Comparing the productivity of the same translator editing with and without highlights (MAIN vs POST), two-thirds

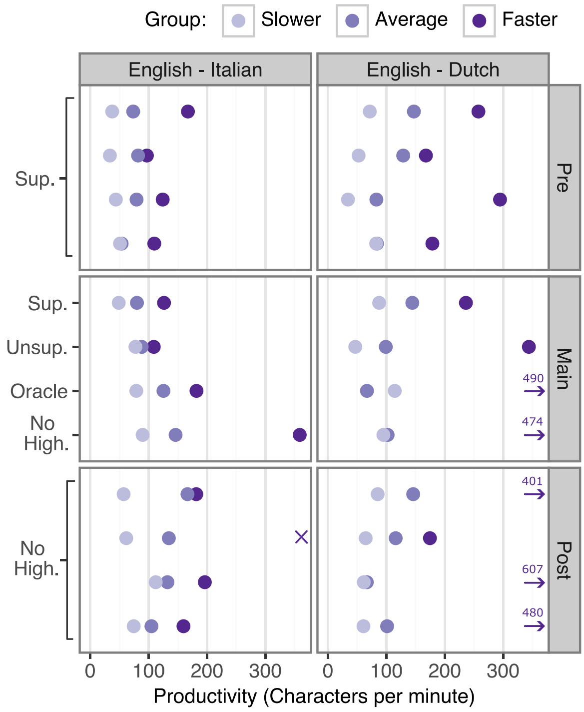


Figure 3: Productivity of post-editors across QE4PE stages (PRE, MAIN, POST). The ➔ marks outstanding entries and ✕ marks missing data. Each row corresponds to the same three translators across all stages.

of the translators editing with highlights were up to two times slower on biomedical texts. However, the same proportion of translators was up to three times faster on social media texts across both directions.

In summary, we find that **highlight modalities are not predictive of edit times on their own**, but translation direction and domain play an important role in determining the effect of highlights on editing productivity. We attribute these results to two main factors, which will remain central in the analysis of the following sections: (1) the different *propensity of translators to act upon highlighted issues* in the two tested directions, and (2) the different *nature of errors highlighted across domains*.

4.2 Highlights and Edits

We then examine how highlights are distributed across modalities and how they influence the editing choices of human post-editors.

Agreement Across Modalities First, we quantify how different modalities agree in terms of highlight distribution and editing. We find that

	Base Freq.		Measured				Projected			
	$P(H)$	$P(E)$	$P(E\|H)$	Λ_E	$P(H\|E)$	Λ_H	$\vec{P}(E\|H)$	$\vec{\Lambda}_E$	$\vec{P}(H\|E)$	$\vec{\Lambda}_H$
English→Italian										
No High.	–	0.05	–	–	–	–	–	–	–	–
Random	0.16	–	–	–	–	–	0.06	1.20	0.18	1.20
Oracle	0.15	0.12	**0.37**	**4.62**	<u>**0.45**</u>	**4.1**	$\mathbf{0.18}_{\downarrow 0.19}$	$\mathbf{6.00}_{\uparrow 1.38}$	<u>$\mathbf{0.55}$</u>$_{\uparrow 0.10}$	$\mathbf{4.23}_{\uparrow 0.14}$
Unsup.	0.16	0.13	0.25	2.27	0.21	2.2	$0.11_{\downarrow 0.14}$	$2.75_{\uparrow 0.48}$	$0.37_{\uparrow 0.16}$	$2.47_{\uparrow 0.26}$
Sup.	0.12	0.16	0.28	2.00	0.22	2.0	$0.14_{\downarrow 0.14}$	$3.50_{\uparrow 1.50}$	$0.35_{\uparrow 0.13}$	$3.18_{\uparrow 1.18}$
English→Dutch										
No High.	–	0.14	–	–	–	–	–	–	–	–
Random	0.17	–	–	–	–	–	0.16	1.14	0.19	1.19
Oracle	0.20	0.10	**0.26**	<u>**4.33**</u>	<u>**0.53**</u>	3.12	$\mathbf{0.28}_{\uparrow 0.02}$	$\mathbf{2.55}_{\downarrow 1.78}$	<u>$\mathbf{0.40}$</u>$_{\downarrow 0.13}$	$2.35_{\downarrow 0.77}$
Unsup.	0.20	0.11	0.20	2.50	0.36	2.00	$0.22_{\uparrow 0.02}$	$1.83_{\downarrow 0.67}$	$0.31_{\downarrow 0.05}$	$1.72_{\downarrow 0.28}$
Sup.	0.12	0.09	0.24	3.43	0.33	**3.30**	$\mathbf{0.28}_{\uparrow 0.04}$	$2.33_{\downarrow 1.10}$	$0.24_{\downarrow 0.09}$	$\mathbf{2.40}_{\downarrow 0.90}$

Table 4: Highlighting (H) and editing (E) average statistics across directions and highlight modalities. **Measured**: actual edits performed in the specified modality. **Projected**: using modality highlights over **No Highlight** edits to account for editing biases (Section 4.2). Random highlights matching average word frequencies are used as **Random** baseline, and Projected increases$_{\uparrow}$/decreases$_{\downarrow}$ compared to Measured counterparts are shown. Significant **Oracle** gains over all other modalities are <u>underlined</u> ($p < 0.05$ with Bonferroni correction).

highlight overlaps across modalities range between 15% and 39% when comparing highlight modalities in a pairwise fashion, with the highest overlap for English→Italian social media and English→Dutch biomedical texts.[7] Despite the relatively low highlight agreement, we find an average agreement of 73% for post-edited characters across modalities. This suggests that edits are generally uniform regardless of highlight modalities and are not necessarily restricted to highlighted spans.[8]

Do Highlights Accurately Identify Potential Issues? Table 4 (Base Freq.) shows raw highlight and edit frequencies across modalities. We observe different trends across the two language pairs: for English→Italian, post-editors working with highlights edit more than twice as much as translators with **No Highlight**, regardless of the highlight modality. On the contrary, for English→Dutch they edit 33% less in the same setting. These results suggest a different attitude towards acting upon highlighted potential issues across the two translation directions, with English→Italian translators appearing to be conditioned to edit more when highlights are present. We introduce four metrics to quantify highlights-edits overlap:

- $P(E|H)$ and $P(H|E)$, reflecting highlights' *precision* and *recall* in predicting edits, respectively.
- $\Lambda_E \stackrel{\text{def}}{=} P(E|H)/P(E|\neg H)$ shows how much more likely an edit is to fall within rather than outside highlighted characters.
- $\Lambda_H \stackrel{\text{def}}{=} P(H|E)/P(H|\neg E)$ shows how much more likely it is for a highlight to mark edited rather than unmodified spans.

Intuitively, character-level recall $P(H|E)$ should be more indicative of highlight quality compared to precision $P(E|H)$, provided that word-level highlights can be useful even when not minimal.[9] Table 4 (Measured) shows metric values across the three highlight modalities (breakdowns by domain and speed shown in Tables 13 and 14). As expected, **Oracle** highlights obtain the best performance in terms of precision and recall, with $P(H|E)$, in particular, being significantly higher than the other two modalities across both directions.

[7] Scores are normalized to account for highlight frequencies across modalities. Agreement is shown in Table 11.

[8] Editing agreement is shown in Figure 7.

[9] For example, if the fully-highlighted word *tradutt<u>ore</u>* is changed to its feminine version *tradutt<u>rice</u>*, $P(H|E) = 1$ (edit correctly and fully predicted) but $P(E|H) = 0.3$ since word stem characters are left unchanged.

Surprisingly, **we find no significant precision and recall differences between Supervised and Unsupervised highlights**, despite the word-level QE training of XCOMET used in the former modality. Moreover, they support the potential of unsupervised, model internals-based techniques to complement or substitute more expensive supervised approaches. Still, likelihood ratios $\Lambda_E, \Lambda_H \gg 1$ for all modalities and directions indicate that highlights are 2–4 times more likely to precisely and comprehensively encompass edits than non-highlighted texts. This suggests that even imperfect highlights that do not reach **Oracle**-level quality might effectively direct editing efforts toward potential issues. We validate these observations by fitting a zero-inflated negative binomial mixed-effects model to predict segment-level edit rates. Results confirm a significantly higher edit rate for English→Italian highlighted modalities and the social media domain with $p < 0.001$ (features and significances shown in Appendix Table 9). We find a significant zero inflation associated with translator identity, suggesting the choice of leaving MT outputs unedited is highly subjective.

Do Highlights Influence Editing Choices? Since in Section 4.1 we found the proportion of highlighted characters to impact the editing rate of translators, we question whether the relatively high $P(E|H)$ and $P(H|E)$ values might be artificially inflated by the eagerness of translators to intervene on highlighted spans. In other words, *do highlights identify actual issues, or do they condition translators to edit when they otherwise would not?* To answer this, we propose to *project* highlights from a selected modality—in which highlights were shown during editing—onto the edits performed by the **No Highlight** translators on the same segments. The resulting difference between measured and projected metrics can then be taken as an estimate for the impact of highlight presentation on their resulting accuracy.

To further ensure the soundness of our analysis, we use a set of projected **Random** highlights as a lower bound for highlight performance. To make the comparison fair, **Random** highlights are created by randomly highlighting words in MT outputs matching the average word-level highlight frequency across all highlighted modalities given the current domain and translation direction. Table 4 (Projected) shows results for the three highlighted modalities. First, all projected metrics remain consistently above the **Random** baseline, suggesting a higher-than-chance ability to identify errors even for worst-performing highlight modalities. Projected precision scores $\overrightarrow{P}(E|H)$ depend on edit frequency, and hence see a major decrease for English→Italian, where the **No Highlight** edit rate $P(E)$ is much lower. However, the increase in $\overrightarrow{\Lambda}_E$ across all English→Italian modalities confirms that, despite the lower edit proportion, highlighted texts remain notably more likely to be edited than non-highlighted ones. Conversely, the lower $\overrightarrow{\Lambda}_E$, $\overrightarrow{P}(H|E)$ and $\overrightarrow{\Lambda}_H$ for English→Dutch show that edits become much less skewed towards highlighted spans in this direction when accounting for presentation bias.

Overall, while the presence of highlights makes English→Italian translators more likely to intervene in MT outputs, their location in the MT output often pinpoints issues that would be edited regardless of highlighting. English→Dutch translators, on the contrary, intervene at roughly the same rate regardless of highlights presence, but their edits are focused mainly on highlighted spans when they are present. This difference is consistent across all subjects in the two directions despite the identical setup and comparable MT and QE quality across languages. This suggests that cultural factors might play a non-trivial role in determining the usability and influence of QE methods regardless of span accuracy, a phenomenon previously observed in human-AI interaction studies (Ge et al., 2024).

4.3 Quality Assessment

We continue our assessment by inspecting the quality of MT and post-edited outputs along three dimensions. First, we use XCOMET segment-level QE ratings as an automatic approximation of quality and compare them to human-annotated quality scores collected in the last phase of our study. For efficiency, these are obtained for the 0–100 Direct Assessment scale commonly used in QE evaluation (Specia et al., 2020), but following an initial step of MQM error annotation to condition scoring on found errors, as prescribed by the ESA protocol (Kocmi et al., 2024b). Then, MQM error span annotations are used to analyze the distribution of error categories. Finally, we manually assess critical errors, which were inserted to quantify highlight modalities effect on unambiguous issues.

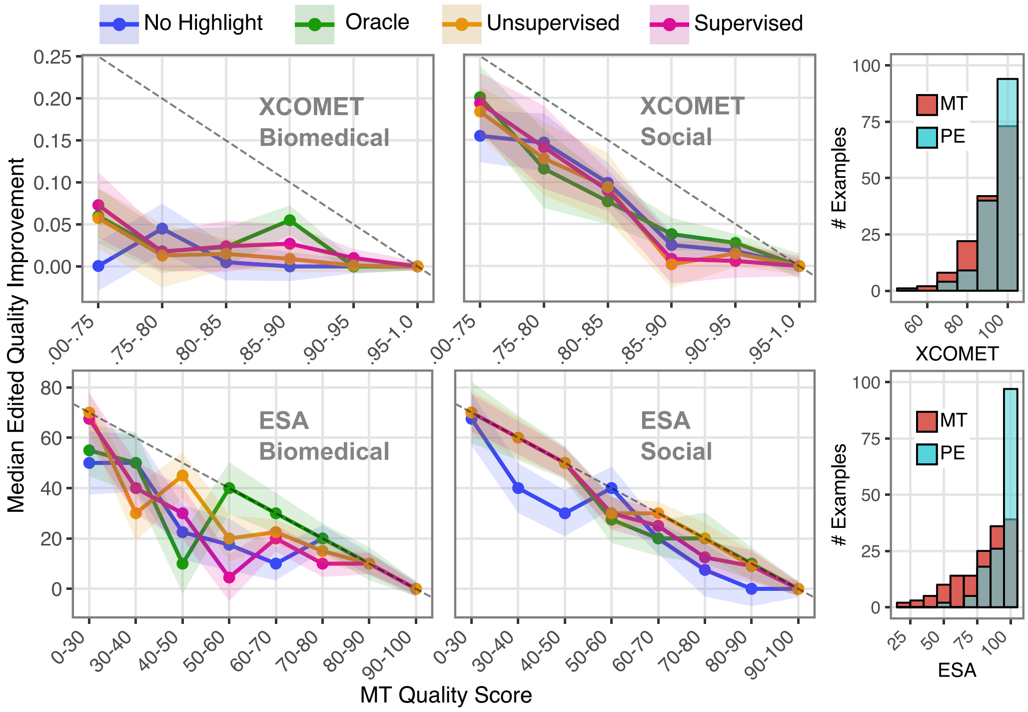


Figure 4: Median quality improvement for post-edited segments at various initial MT quality levels across domains and highlight modalities. Quality scores are estimated using XCOMET segment-level QE (top) and professional ESA annotations (bottom). Histograms show example counts across quality bins for the two metrics. Dotted lines show upper bounds for quality improvements given starting MT quality.

Do Highlights Influence Post-Editing Quality? In this stage, we focus particularly on *edited quality improvements*, i.e., how post-editing the same MT outputs under different highlight conditions influences the resulting quality of translations. We operationalize this assessment using human ratings and automatic metrics to score MT and post-edited translations, using their difference as the effective quality gain after the post-editing stage. Scores for this metric are generally positive, i.e., human post-editing improves quality, and bounded by the maximal achievable quality gain given the initial MT quality. Figure 4 shows median improvement values across quality bins defined from the distribution of initial MT quality scores (shown in histograms), in which all post-edited versions of each MT output appear as separate observations. Positive median scores confirm that post-edits generally lead to quality improvements across all tested settings. However, we observe different trends across the two metrics: Across both domains, XCOMET greatly underestimates the human-assessed ESA quality improvement, especially for biomedical texts where it shows negligible improvement regardless of the initial MT quality. These results echorecent findings cautioning users against the poor performance of trained MT metrics for unseen domains and high-quality translations (Agrawal et al., 2024; Zouhar et al., 2024). Focusing on the more reliable ESA scores, we observe large quality improvements from post-editing, as shown by near-maximal quality gains across most bins and highlight modalities. While **No Highlight** seems to underperform other modalities in the social media domain, the lack of more notable differences in gains across highlight modalities suggests that **highlights' quality impact might not be evident in terms of segment-level quality**, motivating our next steps in the quality analysis.

We also find no clear relationship between translator speed and edited quality improvements, suggesting that higher productivity does not come at a cost for faster translators (Figure 10). This finding confirms that neglecting errors is not the cause of different editing patterns from previous sections.

Which Error Types Do Highlights Identify? Table 5 shows a breakdown of MQM annotations for MT and all highlight modalities using the *Accuracy*, *Style* and *Linguistic* macro-categories

		MT	No High.	Oracle	Unsup.	Sup.
Italian	Acc.	26 / 31	12 / 17	**6 / 12**	30 / 22	22 / 24
	Style	17 / 33	5 / 33	**0 / 15**	5 / 35	4 / 31
	Ling.	12 / 31	2 / 29	**0 / 11**	7 / 20	2 / 16
	Tot.	55 / 95	19 / 79	**6 / 38**	42 / 77	28 / 71
Dutch	Acc.	32 / 40	20 / **31**	26 / 37	**14** / 39	20 / 39
	Style	3 / 29	**1** / 28	**1** / 23	2 / **18**	7 / 50
	Ling.	4 / 26	3 / 18	5 / 28	**2 / 9**	3 / 15
	Tot.	39 / 95	24 / 77	32 / 88	**18 / 66**	30 / 104

Table 5: Minor / major MQM error counts averaged across $n = 3$ translators per highlight modality for every translation direction on the QA MAIN subset. Lowest minor / major error counts per language are **bolded**.

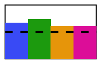
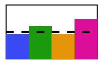
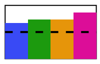
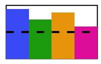
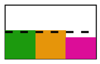
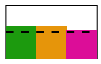
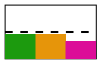
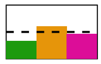
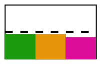
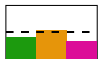
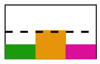
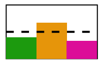
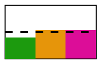
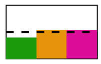
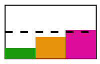
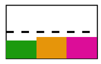
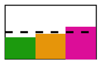
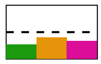
Question	**Italian**	**Dutch**
MT outputs were generally of high quality.		
Provided texts were challenging to translate.		
Highlights ...		
... were generally accurate in detecting potential issues.		
... were generally useful during editing.		
... improved my editing productivity.		
... improved the quality of my translations.		
... required additional editing effort on my part.		
... influenced my editing choices.		
... helped identify errors I'd have otherwise missed.		

Table 6: Post-task questionnaire responses. Bars represent responses ranging from 1–Strongly disagree (no bar) to 5–Strongly agree (full bar), averaged across $n = 3$ translators per language for **No Highlight**, **Oracle**, **Unsupervised**, and **Supervised**. Dotted line mark avg. judgments of 3–Neither agree nor disagree.

of MQM errors.[10] At this granularity, differences across modalities become visible, with overall error counts showing a clear relation to $\vec{\Lambda}_E$ from Table 4 (**Oracle** being remarkably better for English→Italian, with milder and more uniform trends in English→Dutch). At least for English→Italian, these results confirm that an observable quality improvement from editing with highlights is present in the best-case **Oracle** scenario. By contrast, for English→Dutch the **Unsupervised** method is found to outperform even the **Oracle** setting in reducing the amount of errors, while it fares relatively poorly for English→Italian. We also note a different distribution of Accuracy and Style errors, with the former being more common in biomedical texts while the latter appearing more often for translated social media posts (Figure 9). We posit that differences in error types across domains might explain the opposite productivity trends observed in Section 4.1: While highlighted accuracy errors might lead to time-consuming terminology verification in biomedical texts, style errors might be corrected more quickly and naturally in the social media domain.

Do Highlights Detect Critical Errors? We examine whether the critical errors we inserted were detected by different modalities, finding that while most modalities fare decently with more than 62% of critical errors highlighted, **Unsupervised** is the only setting for which all errors are correctly highlighted across both directions. Then, critical errors are manually verified in all outputs, finding that 16–20% more critical errors are edited in highlighted modalities compared to **No Highlight** (full results in Table 12). Hence, **highlights might lead to narrow but tangible quality improvements that can go undetected in coarse quality assessments**, and finer-grained evaluations might be needed to quantify future improvements in word-level QE.

4.4 Usability

In post-task questionnaire answers (Table 6), most translators stated that MT outputs had average-to-high quality and that provided texts were challenging to translate. Highlights were generally found decently accurate, but they were generally not found useful to improve either productivity or quality (including **Oracle** ones). Interestingly, despite the convincing gains for critical errors measured in the last section, most translators stated that highlights did not influence their editing and did not help them identify errors that would have otherwise been missed. Concretely, this suggests that the potential quality improvements might not be easily perceived by

[10] Full micro-category breakdown in Table 12, per-domain breakdown in Figure 9. Category descriptions in Figure 5.

Doc ID - Seg. ID	Source text	Target text	Proposed correction	Error Annotation: Description	Category	Severity	Score
9-1	Specifying peri- and postnatal factors in children born very preterm (VPT) that affect later outcome helps to improve long-term treatment.	Specificare i fattori peri- e post-natali nei bambini nati molto pretermine (VPT) che influenzano il risultato successivo aiuta a migliorare il trattamento a lungo termine.	Specificare i fattori peri- e postnatali nei bambini nati molto pretermine (VPT, Very Preterm) che influenzano il risultato successivo aiuta a migliorare il trattamento a lungo termine.	When we have a foreign acronym, the usual rule is to indicate also the whole term the first time it appears.	Readability	Minor	90
9-2	To enhance the predictability of 5-year cognitive outcome by perinatal, 2-year developmental and socio-economic data.	Migliorare la prevedibilità del risultato cognitivo a 5 anni mediante dati perinatali, di sviluppo e socioeconomici a 2 anni.					100
9-3	5-year infants born VPT were compared to 34 term controls.	I neonati di 5 anni nati VPT sono stati confrontati con 34 nati a termine come controllo.	I neonati di 5 anni nati VPT sono stati confrontati con 34 controlli a termine.		Mistranslation	Minor	70
9-4	The IQ of 5-year infants born VPT was 10 points lower than that of term controls and influenced independently by preterm birth and SES.	Il QI dei bambini di 5 anni nati VPT era di 10 punti inferiore a quello dei nati a termine di controllo, e influenzato indipendentemente dalla nascita pretermine e dai dati SES.	Il QI dei bambini di 5 anni nati VPT era di 10 punti inferiore a quello dei nati a termine e influenzato indipendentemente dalla nascita pretermine e dallo stato socioeconomico (SES).		Mistranslation	Minor	70
		Il QI dei bambini di 5 anni nati VPT era di 10 punti inferiore a quello dei nati a termine di controllo, e influenzato indipendentemente dalla nascita pretermine e dai dati SES.	Il QI dei bambini di 5 anni nati VPT era di 10 punti inferiore a quello dei nati a termine e influenzato indipendentemente dalla nascita pretermine e dallo stato socioeconomico (SES).	Unexplained acronym. Non-expert people could have trouble understanding the meaning.	Untranslated	Minor	
52-1	But with less than 3 months to go for that, I feel I'm not ready yet, but having never taken it, I have nothing to compare it to besides colleagues' advice.	Ma con meno di 3 mesi per farlo, sento di non essere ancora pronto, ma non l'ho mai preso, non ho nulla con cui confrontarlo oltre ai consigli dei colleghi.	Ma con meno di 3 mesi per farlo, sento di non essere ancora pronto, e non avendolo mai fatto, non ho nulla con cui confrontarlo oltre ai consigli dei colleghi.		Mistranslation	Major	30
52-2	Without knowing what I know, they can't know if I'm actually ready yet, but many of them are pushing me to sign up for it.	Senza sapere quello che so, non possono sapere se sono ancora pronta, ma molti di loro mi stanno spingendo a iscrivermi.	Se non hanno idea di quanto sappia, non possono sapere se sono davvero pronta, ma molti di loro mi stanno spingendo a iscrivermi.		Readability	Minor	60
		Senza sapere quello che so, non possono sapere se sono ancora pronta, ma molti di loro mi stanno spingendo a iscrivermi.	Se non hanno idea di quanto sappia, non possono sapere se sono davvero pronta, ma molti di loro mi stanno spingendo a iscrivermi.		Mistranslation	Minor	
52-3	I'm close... I just don't know if I'm 2 months study close.	Ci sono quasi... solo che non so se ce la farò in soli 2 mesi, ma penso di potercela fare.	Ci sono quasi... solo che non so se ce la farò in soli 2 mesi.		Addition	Major	20

Error category	Subcategory	Description
Accuracy Incorrect meaning has been transferred to the source text.	**Addition**	Translation includes the information that is not present in the source and it changes or distorts the original message.
	Omission	Translation is missing the information that is present in the source, which is important to convey the message.
	Mistranslation	Translation does not accurately represent the source content meaning.
	Inconsistency	There are internal inconsistencies in the translation (for example, using different verb forms in the bullet list or in CTAs, calling the same UI element differently, terminology used inconsistently etc).
	Untranslated	Content that should have been translated has been left untranslated.
Linguistic Official linguistic reference sources such as grammar books.	**Punctuation**	Punctuation is used incorrectly (for the locale or style), including missing or extra white spaces and the incorrect use of space (non-breaking space). Violation of typographic conventions of the locale.
	Spelling	Issues related to spelling of words, including typos, wrong word hyphenation, word breaks and capitalization.
	Grammar	Issues related to the grammar or syntax of the text, other than spelling.
Style Not suitable/native; too literal or awkward.	**Inconsistent Style**	Style is inconsistent within a text.
	Readability	Translation does not read well (due to heavy sentence structure, frequent repetitions, unidiomatic).
	Wrong Register	Inappropriate style for the specific subject field, the level of formality, and the mode of discourse (e.g., written text versus transcribed speech).

Severity level	Description
Major	The Severity Level of an error that seriously affects the understandability, reliability, or usability of the content for its intended purpose or hinders the proper use of the product or service due to a significant loss or change in meaning or because the error appears in a highly visible or important part of the content.
Minor	The Severity Level of an error that does not seriously impede the usability, understandability, or reliability of the content for its intended purpose, but has a limited impact on, for example, accuracy, stylistic quality, consistency, fluency, clarity, or general appeal of the content.
Neutral	The Severity Level of an error that differs from a quality evaluator's preferential translation or that is flagged for the translator's attention but is an acceptable translation.

Figure 5: **Top:**QA interface with cropped examples of biomedical and social media texts with error annotations (Biomedical: post-edited segments with **No Highlight**; Social media: MT outputs). **Bottom:** Annotations instructions for our MQM-inspired error taxonomy.

translators and might have secondary importance compared to the extra cognitive load elicited by highlighted spans. When asked to comment about highlights, several translators called them *''more of an eye distraction, as they often weren't actual mistakes''* and *''not quite accurate enough to rely on them as a suggestion''*. Some translators also stated that missed errors led them to *''disregarding the highlights to focus on checking each sentence''*. Despite their high quality, only one editor working with **Oracle** highlights found highlights helpful in *''making the editing process faster and somehow easier''*. Taken together, these comments convincingly point to a negative perception of the quality and usefulness of highlights, suggesting that **improvement in QE accuracy may not be sufficient to improve QE usefulness** in editors' eyes.

5 Conclusion

This study evaluated the impact of various error-span highlighting modalities, including automatic and human-made ones, on the productivity and quality of human post-editing in a realistic professional setting. Our findings highlight the importance of domain, language, and editors' speed in determining highlights' effect on productivity and quality, underscoring the need for broad evaluations encompassing diverse settings. The limited gains of human-made highlights over automatic QE and their indistinguishable perception from editors' assessment indicate that further gains in the accuracy of these techniques might not be the determining factor in improving their integration into post-editing workflows. In particular, future work might explore other directions to further assess and improve the usability of word-level QE highlights, for example, studying their impact on non-professional translators and language learners or combining them with edit suggestions to justify the presence of error spans.

6 Limitations

Our study presents certain limitations that warrant consideration when interpreting its findings and for guiding future research.

Firstly, while we included two domains and translation directions to improve the generalizability of our findings, our results suggest that language and domain play an important role in defining the effectiveness of word-level QE for human post-editing. While we observed mild gains from word-level QE on our tested mid-resourced translation directions (English→Italian and English→Dutch), we expect limited, if any, benefit of such approaches in low-resource languages and domains for which MT systems and QE methods are likely to underperform (Sarti et al., 2022; Zouhar et al., 2024). Furthermore, the domains tested in our study (biomedical and social media posts) provided concrete challenges in the form of specialized terminology and idiomatic expressions, respectively, which are known to hinder the quality of MT outputs (Neves et al., 2024; Bawden and Sagot, 2023). While future work should ensure our findings can be extended to other domains and languages, the limited benefits brought by the tested word-level QE methods in challenging settings suggest a limited usefulness for higher-resource languages and more standard domains such as news or Wiki texts.

Secondly, we acknowledge that several design choices in our evaluation setup, rather than pertaining to the QE methods themselves, may have influenced our results. These include, for instance, the specific procedure for discretizing continuous scores from the **Unsupervised** method into error spans, and the method of obtaining ***oracle*** highlights via majority voting among post-editors. While we believe these choices are justified within the context of our study, their impact on the outcomes cannot be entirely discounted. Future studies might benefit from a more fine-grained assessment of how such low-level decisions influence the perceived accuracy and usability of word-level QE.

Finally, subjective factors such as the translators' inherent propensity to edit, their prior opinions on the role of MT in post-editing, and their individual editing styles inevitably influenced both quantitative and qualitative assessments in this study. Although we attempted to mitigate these effects by ensuring a controlled evaluation setup for all professional translators and by using averaged judgments for translators working on the same highlight modality, we acknowledge that subjectivity might limit the reproducibility of our findings.

7 Broader Impact and Ethical Considerations

Our study explicitly centers the experience of professional translators, responding to recent calls for user-centered evaluation of translation technologies. By prioritizing translators' perspectives and productivity, we aim to contribute to methods that complement rather than replace human expertise. Our findings highlight a gap between user perception and measured quality improvements, suggesting that future efforts should focus primarily on improving the usability of these methods in editing interfaces. In particular, new assistive approaches for post-editing should not only strive to increase productivity but rather reduce the cognitive burden associated with post-editing work. This insight is crucial for designing more user-centered quality estimation tools that genuinely support human work. Ultimately, our results suggest that subjective norms across different domains and cultures play an important role in determining the effectiveness of proposed methodologies, underscoring the importance of accounting for human factors when designing such evaluations. All participants in this study were professional translators who provided informed consent. The research protocol ensured anonymity and voluntary participation, with translators recruited and remunerated through professional translation providers. The released materials further promote transparency, enabling other researchers to reproduce and build upon our findings.

Acknowledgments

The authors acknowledge the support of Imminent, which funded the data collection through a Language Technology Grant. The authors are grateful to Translated S.r.l. and Global Textware B.v. and to all translators involved in this study for agreeing to participate in the data collection process. We thank TACL reviewers and editors for their insightful comments. Gabriele Sarti, Grzegorz Chrupała and Arianna Bisazza also acknowledge the support of the Dutch Research Council (NWO) for the project InDeep (NWA.1292.19.399). Arianna Bisazza is supported by the NWO Talent Program (VI.Vidi.221C.009). Ana Guerberof Arenas is supported by the ERC Grant CoG INCREC 101086819.

References

Abubakar Abid, Ali Abdalla, Ali Abid, Dawood Khan, Abdulrahman Alfozan, and James Zou. 2019. Gradio: Hassle-free sharing and testing of ML models in the wild. *CoRR*, cs.LG/1906.02569v1.

Sweta Agrawal, António Farinhas, Ricardo Rei, and Andre Martins. 2024. Can automatic metrics assess high-quality translations? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14491–14502. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.802

Chantal Amrhein, Nikita Moghe, and Liane Guillou. 2022. ACES: Translation accuracy challenge sets for evaluating machine translation metrics. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 479–513. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.wmt-1.57

Chantal Amrhein, Nikita Moghe, and Liane Guillou. 2023. ACES: Translation accuracy challenge sets at WMT 2023. In *Proceedings of the Eighth Conference on Machine Translation*, pages 695–712. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.wmt-1.57

Rachel Bawden and Benoît Sagot. 2023. RoCS-MT: Robustness challenge set for machine translation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 198–216. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.wmt-1.21

Frederic Blain, Chrysoula Zerva, Ricardo Rei, Nuno M. Guerreiro, Diptesh Kanojia, José G. C. de Souza, Beatriz Silva, Tânia Vaz, Yan Jingxuan, Fatemeh Azadi, Constantin Orasan, and André Martins. 2023. Findings of the WMT 2023 shared task on quality estimation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 629–653. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.wmt-1.52

John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. 2004. Confidence estimation for machine translation. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 315–321.

COLING. https://doi.org/10.3115/1220355.1220401

Sven Coppers, Jan Van den Bergh, Kris Luyten, Karin Coninx, Iulianna Van der Lek-Ciudin, Tom Vanallemeersch, and Vincent Vandeghinste. 2018. Intellingo: An intelligible translation environment. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–13. https://doi.org/10.1145.3173574.3174098

David Dale, Elena Voita, Loic Barrault, and Marta R. Costa-jussà. 2023. Detecting and mitigating hallucinations in machine translation: Model internal workings alone do well, sentence similarity even better. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 36–50. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.3

Sugyeong Eo, Chanjun Park, Hyeonseok Moon, Jaehyung Seo, and Heuiseok Lim. 2022. Word-level quality estimation for Korean-English neural machine translation. *IEEE Access*, 10:44964–44973. https://doi.org/10.1109/ACCESS.2022.3169155

Johannes Eschbach-Dymanus, Frank Essenberger, Bianka Buschbeck, and Miriam Exel. 2024. Exploring the effectiveness of LLM domain adaptation for business IT machine translation. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 610–622. European Association for Machine Translation (EAMT).

Patrick Fernandes, Daniel Deutsch, Mara Finkelstein, Parker Riley, André Martins, Graham Neubig, Ankush Garg, Jonathan Clark, Markus Freitag, and Orhan Firat. 2023. The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1066–1083. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.wmt-1.100

Marina Fomicheva, Piyawat Lertvittayakumjorn, Wei Zhao, Steffen Eger, and Yang Gao. 2021. The Eval4NLP shared task on explainable quality estimation: Overview and results. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 165–178. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.eval4nlp-1.17

Marina Fomicheva and Lucia Specia. 2019. Taking MT evaluation metrics to extremes: Beyond correlation with human judgments. *Computational Linguistics*, 45(3):515–558. https://doi.org/10.1162/coli_a_00356

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. 2020. Unsupervised quality estimation for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:539–555. https://doi.org/10.1162/tacl_a_00330

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. https://doi.org/10.1162/tacl_a_00437

Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. 2024. Are LLMs breaking MT metrics? Results of the WMT24 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 47–81. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.wmt-1.2

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1050–1059, New York, New York, USA. PMLR.

Xiao Ge, Chunchen Xu, Daigo Misaki, Hazel Rose Markus, and Jeanne L. Tsai. 2024. How culture shapes what people want from AI. In *Proceedings of the 2024 CHI Conference* *on Human Factors in Computing Systems*, CHI '24. New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/3613904.3642660

Naman Goyal, Jingfei Du, Myle Ott, Giri Anantharaman, and Alexis Conneau. 2021. Larger-scale transformers for multilingual masked language modeling. *CoRR*, cs.CL/2105.00572v1. https://doi.org/10.18653/v1/2021.repl4nlp-1.4

Ana Guerberof-Arenas and Joss Moorkens. 2023. Ethics and machine translation: The end user perspective. In Helena Moniz and Carla Parra Escartín, editors, *Towards Responsible Machine Translation: Ethical and Legal Considerations in Machine Translation*, pages 113–133. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-031-14689-3_7

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995. https://doi.org/10.1162/tacl_a_00683

Nico Herbig, Tim Düwel, Santanu Pal, Kalliopi Meladaki, Mahsa Monshizadeh, Antonio Krüger, and Josef van Genabith. 2020. MMPE: A multi-modal interface for post-editing machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1691–1702. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.155

Anas Himmi, Guillaume Staerman, Marine Picot, Pierre Colombo, and Nuno M. Guerreiro. 2024. Enhanced hallucination detection in neural machine translation through simple detector aggregation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18573–18583. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.1033

Fabio Kepler, Jonay Trénous, Marcos Treviso, Miguel Vera, and André F. T. Martins. 2019. OpenKiwi: An open source framework for quality estimation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 117–122. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-3020

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Makoto Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović, and Mariya Shmatova. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.wmt-1.1

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinthór Steingrímsson, and Vilém Zouhar. 2024a. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.wmt-1.1

Tom Kocmi and Christian Federmann. 2023. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.wmt-1.64

Tom Kocmi, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popović, Mrinmaya Sachan, and Mariya Shmatova. 2024b. Error span annotation: A balanced approach for human evaluation of machine translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1440–1453. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.wmt-1.131

Hans P. Krings. 2001. *Repairing texts: Empirical investigations of machine translation post-editing processes*. Kent, Ohio and London. Kent State University Press.

Christoph Leiter, Piyawat Lertvittayakumjorn, Marina Fomicheva, Wei Zhao, Yang Gao, and Steffen Eger. 2024. Towards explainable evaluation metrics for machine translation. *Journal of Machine Learning Research*, 25(75):1–49.

Haijun Li, Tianqi Shi, Zifu Shang, Yuxuan Han, Xueyu Zhao, Hao Wang, Yu Qian, Zhiqiang Qian, Linlong Xu, Minghao Wu, Chenyang Lyu, Longyue Wang, Gongbo Tang, Weihua Luo, Zhao Xu, and Kaifu Zhang. 2025. Transbench: Benchmarking machine translation for industrial-scale applications. *CoRR*, cs.CL/2505.14244v1.

Zheng Wei Lim, Ekaterina Vylomova, Charles Kemp, and Trevor Cohn. 2024. Predicting human translation difficulty with neural machine translation. *Transactions of the Association for Computational Linguistics*, 12:1479–1496. https://doi.org/10.1162/tacl_a_00714

Zhongtao Liu, Parker Riley, Daniel Deutsch, Alison Lui, Mengmeng Niu, Apurva Shah, and Markus Freitag. 2024. Beyond human-only: Evaluating human-machine collaboration for collecting high-quality translation data. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1095–1106. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.wmt-1.110

Arle Richard Lommel, Aljoscha Burchardt, and Hans Uszkoreit. 2013. Multidimensional quality metrics: A flexible system for assessing translation quality. In *Proceedings of Translating and the Computer 35*. London, UK. Aslib.

Samuel Läubli, Patrick Simianer, Joern Wuebker, Geza Kovacs, Rico Sennrich, and Spence Green. 2022. The impact of text presentation on translator performance. *Target. International Journal of Translation Studies*, 34(2):309–342. https://doi.org/10.1075/target.20006.lau

Nikita Mehandru, Sweta Agrawal, Yimin Xiao, Ge Gao, Elaine Khoong, Marine Carpuat, and Niloufar Salehi. 2023. Physician detection of clinical harm in machine translation: Quality estimation aids in reliance and back-translation identifies critical errors. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11633–11647. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.712

Yasmin Moslem, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023. Adaptive machine translation with large language models. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 227–237. European Association for Machine Translation. https://doi.org/10.18653/v1/2023.wmt-1.82

Mariana Neves, Cristian Grozea, Philippe Thomas, Roland Roller, Rachel Bawden, Aurélie Névéol, Steffen Castle, Vanessa Bonato, Giorgio Maria Di Nunzio, Federica Vezzani, Maika Vicente Navarro, Lana Yeganova, and Antonio Jimeno Yepes. 2024. Findings of the WMT 2024 biomedical translation shared task: Test sets on abstract level. In *Proceedings of the Ninth Conference on Machine Translation*, pages 124–138. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.wmt-1.6

Mariana Neves, Antonio Jimeno Yepes, Aurélie Névéol, Rachel Bawden, Giorgio Maria Di Nunzio, Roland Roller, Philippe Thomas, Federica Vezzani, Maika Vicente Navarro, Lana Yeganova, Dina Wiemann, and Cristian Grozea. 2023. Findings of the WMT 2023 biomedical translation shared task: Evaluation of ChatGPT 3.5 as a comparison system. In *Proceedings of the Eighth Conference on Machine Translation*, pages 43–54. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.wmt-1.2

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846. https://doi.org/10.1038/s41586-024-07335-x, PubMed: 38839963

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C. Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: UNBabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585. Association for Computational Linguistics.

Ricardo Rei, Ana C. Farinha, Chrysoula Zerva, Daan van Stigt, Craig Stewart, Pedro Ramos, Taisiya Glushkova, André F. T. Martins, and Alon Lavie. 2021. Are references really needed? UNBabel-IST 2021 submission for the metrics shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1030–1040. 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. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.213

Gabriele Sarti, Arianna Bisazza, Ana Guerberof-Arenas, and Antonio Toral. 2022. DivEMT: Neural machine translation post-editing effort across typologically diverse languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7795–7816. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-main.532

Beatrice Savoldi, Alan Ramponi, Matteo Negri, and Luisa Bentivogli. 2025. Translation in the hands of many: Centering lay users in machine translation interactions. *CoRR*, cs.CL/2502.13780v1.

Raksha Shenoy, Nico Herbig, Antonio Krüger, and Josef van Genabith. 2021. Investigating the helpfulness of word-level quality estimation for post-editing machine translation output. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10173–10185. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.799

Lucia Specia, Frédéric Blain, Marina Fomicheva, Erick Fonseca, Vishrav Chaudhary, Francisco Guzmán, and André F. T. Martins. 2020. Findings of the WMT 2020 shared task on quality estimation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 743–764. Association for Computational Linguistics.

Lucia Specia, Marco Turchi, Nicola Cancedda, Nello Cristianini, and Marc Dymetman. 2009. Estimating the sentence-level quality of machine translation systems. In *Proceedings of the 13th Annual Conference of the European Association for Machine Translation*. European Association for Machine Translation.

Jiao Sun, Swabha Swayamdipta, Jonathan May, and Xuezhe Ma. 2022. Investigating the benefits of free-form rationales. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5867–5882. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.findings-emnlp.432

Aleš Tamchyna. 2021. Deploying MT quality estimation on a large scale: Lessons learned and open questions. In *Proceedings of Machine Translation Summit XVIII: Users and Providers Track*, pages 291–305. 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.

Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-acl.304

Brian Thompson and Matt Post. 2020. Automatic machine translation evaluation in many languages via zero-shot paraphrasing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 90–121. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.8

Marco Turchi, Antonios Anastasopoulos, José G. C. de Souza, and Matteo Negri. 2014. Adaptive quality estimation for machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 710–720. Association for Computational Linguistics. https://doi.org/10.3115/v1/P14-1067

Marco Turchi, Matteo Negri, and Marcello Federico. 2013. Coping with the subjectivity of human judgements in MT quality estimation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 240–251. Association for Computational Linguistics.

Helena Vasconcelos, Gagan Bansal, Adam Fourney, Q. Vera Liao, and Jennifer Wortman Vaughan. 2025. Generation probabilities are not enough: Uncertainty highlighting in AI code completions. *ACM Transactions on Computer-Human Interaction*, 32(1). https://doi.org/10.1145/3702320

Weijia Xu, Sweta Agrawal, Eleftheria Briakou, Marianna J. Martindale, and Marine Carpuat. 2023. Understanding and detecting hallucinations in neural machine translation via model introspection. *Transactions of the Association for Computational Linguistics*, 11:546–564. https://doi.org/10.1162/tacl_a_00563

Chrysoula Zerva, Frederic Blain, José G. C. De Souza, Diptesh Kanojia, Sourabh Deoghare, Nuno M. Guerreiro, Giuseppe Attanasio, Ricardo Rei, Constantin Orasan, Matteo Negri, Marco Turchi, Rajen Chatterjee, Pushpak Bhattacharyya, Markus Freitag, and André Martins. 2024. Findings of the quality estimation shared task at WMT 2024: Are LLMs closing the gap in QE? In *Proceedings of the Ninth Conference on Machine Translation*, pages 82–109. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.wmt-1.3

Chrysoula Zerva, Frédéric Blain, Ricardo Rei, Piyawat Lertvittayakumjorn, José G. C. de Souza, Steffen Eger, Diptesh Kanojia, Duarte Alves, Constantin Orăsan, Marina Fomicheva, André F. T. Martins, and Lucia Specia. 2022. Findings of the WMT 2022 shared task on quality estimation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 69–99. Association for Computational Linguistics.

Vilém Zouhar, Shuoyang Ding, Anna Currey, Tatyana Badeka, Jenyuan Wang, and Brian Thompson. 2024. Fine-tuned machine translation metrics struggle in unseen domains. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 488–500. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-short.45

Vilém Zouhar, Tom Kocmi, and Mrinmaya Sachan. 2025. AI-assisted human evaluation of machine translation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4936–4950, Albuquerque, New Mexico. Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.naacl-long.255

Vilém Zouhar, Michal Novák, Matúš Žilinec, Ondřej Bojar, Mateo Obregón, Robin L. Hill, Frédéric Blain, Marina Fomicheva, Lucia Specia, and Lisa Yankovskaya. 2021a. Backtranslation feedback improves user confidence in MT, not quality. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 151–161. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.14

Vilém Zouhar, Martin Popel, Ondřej Bojar, and Aleš Tamchyna. 2021b. Neural machine translation quality and post-editing performance. In

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 10204–10214. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.801

A Filtering Details for QE4PE Data

1. *Documents should contain between 4 and 10 segments, each containing 10–100 words (959 docs).* This ensures that all documents are roughly uniform in terms of size and complexity to maintain a steady editing flow (Section 3.5).

2. *The average segment-level QE score predicted by XCOMET-XXL is between 0.3 and 0.95, with no segment below 0.3 (429 docs).* This forces segments to have a decent but still imperfect quality, excluding fully wrong translations.

3. *At least 3 and at most 20 errors spans per document, with no more than 30% of words in the document being highlighted (351 docs).* This avoids overwhelming the editor with excessive highlighting, while still ensuring error presence.

The same heuristics were applied to both translation directions, selecting only documents matching our criteria in both cases.

Remove negation (13-6)	
English	No significant differences were found with respect to principal diagnoses [...]
Dutch	Er werden geen significante verschillen → significante verschillen gevonden met betrekking tot de belangrijkste diagnoses [...]
Title literal translation (16-3)	
English	The Last of Us is an easy and canonical example of dad-ification. [...]
Italian	The Last of Us → L'ultimo di noi è un esempio facile e canonico di dad-ification. [...]
Wrong term (48-5)	
English	[...], , except for alkaline phosphatase.
Italian	[...], ad eccezione della fosfatasi alcalina → chinasi proteica.

Table 7: Examples of original → manually inserted critical errors with document-segment ID from Table 12.

Target: Seg. Edit Time, 5s bins from 0 to 600s		
Feature	**Coeff.**	**Significance**
(Intercept)	1.67	***
MT Num. Chars	2.42	***
Highlight Ratio %	1.59	***
Target Lang.: ITA	−0.34	***
Text Domain: Social	0.31	***
Oracle Highlight	−0.79	.
Sup. Highlight	0.02	
Unsup. Highlight	−0.07	
MT XCOMET QE Score	0.01	***
ITA:**Oracle**	0.91	***
ITA:**Sup.**	1.18	***
ITA:**Unsup.**	0.48	***
Social:**Oracle**	−0.19	**
Social:**Sup.**	−0.34	***
Social:**Unsup.**	−0.22	***
Highlight Ratio:**Oracle**	−0.83	*
Highlight Ratio:**Sup.**	−1.33	***
Edit Order Translator ID Segment ID	**Random Factors**	

Table 8: Details for the negative binomial mixed-effect model used for the productivity analysis of Section 4.1.

Target: % of edited characters in a segment (0–100).		
Feature	**Coeff.**	**Significance**
(Intercept)	21.0	***
MT Num. Chars	10.3	***
Highlight Ratio %	7.1	***
Target Lang.: ITA	−9.9	***
Text Domain: Social	10.9	***
Oracle Highlight	−5.2	
Sup. Highlight	−4.7	
Unsup. Highlight	−0.9	
ITA:**Oracle**	12.2	***
ITA:**Sup.**	15.9	***
ITA:**Unsup.**	13.4	***
Social:**Oracle**	3.5	***
Social:**Sup.**	−0.4	
Social:**Unsup.**	2.1	**
Highlight Ratio:**Oracle**	−0.18	
Highlight Ratio:**Sup.**	−1.78	***
Edit Order Translator ID Segment ID	**Random Factors**	
MT Num. Chars Target Lang Text Domain Translator ID	**Zero-Inflation Factors**	

Table 9: Details for the zero-inflated negative binomial mixed-effect model used for the editing analysis of Section 4.2. The model achieves an RMSE of 0.11 and an R^2 of 0.98.

Identifier	Job	Eng. Lvl	Trans. YoE	Post-edit YoE	Post-edit %	Adv. CAT YoE	MT good/bad for:	Post-edit comment
eng-ita-nohigh-fast	Freelance (FT)	C1	2–5	2–5	100%	Often	Good: Productivity, quality, repetitive work.	PE better than from scratch when consistency is needed.
eng-ita-nohigh-avg	Freelance (PT)	C1	>10	<2	20%	Often	Good: Productivity, repetitive work. Bad: less creative.	PE produces unnatural sentences.
eng-ita-nohigh-slow	Freelance (PT)	C2	>10	2–5	40%	Sometimes	Good: creativity.	Good for time saving.
eng-ita-oracle-fast	Freelance (FT)	C2	5–10	2–5	60%	Sometimes	Good: Productivity, repetitive work. Bad: less creative.	Good for productivity, humans always needed.
eng-ita-oracle-avg	Freelance (FT)	C2	5–10	5–10	20%	Always	Good: productivity, terminology.	Good for tech docs, not for articulated texts.
eng-ita-oracle-slow	Freelance (FT)	C2	2–5	5–10	80%	Always	Good: Productivity, repetitive work.	Useful for consistency and productivity, unless creativity is needed.
eng-ita-unsup-fast	Freelance (FT)	C1	<2	<2	60%	Often	Good: Productivity, terminology. Bad: less creative.	Humans will always be needed in translation.
eng-ita-unsup-avg	Freelance (FT)	C1	>10	2–5	60%	Often	Good: Productivity, repetitive work. Bad: less creative.	An opportunity for translators.
eng-ita-unsup-slow	Freelance (FT)	C1	5–10	5–10	80%	Always	Good: Productivity, repetitive work. Bad: less creative.	Good for focusing on detailed/cultural/creative aspects of translations.
eng-ita-sup-fast	Freelance (PT)	C1	>10	2–5	40%	Often	Good: Productivity, quality, repetitive work, terminology.	Improves quality and consistency.
eng-ita-sup-avg	Freelance (FT)	C1	>10	5–10	100%	Always	Good: Productivity, repetitive work. Bad: less creative.	Consistency improved, but less variance means less creativity.
eng-ita-sup-slow	Freelance (FT)	C1	>10	2–5	20%	Always	Good: Productivity, creativity, quality, repetitive work.	Good for productivity, but does not work on creative texts.
eng-nld-nohigh-fast	Freelance (FT)	C1	>10	>10	40%	Often	Good: Productivity, terminology. Bad: creativity.	Widespread but still too literal
eng-nld-nohigh-avg	Freelance (FT)	C2	>10	2–5	40%	Always	Good: Repetitive work. Bad: creativity, often wrong, worse quality.	Increase in productivity to save on costs brings down quality.
eng-nld-nohigh-slow	Freelance (FT)	C2	>10	5–10	100%	Often	Good: Creativity, quality, repetitive work, terminology.	Working with MT can be creative beyond PE.
eng-nld-oracle-fast	Freelance (FT)	C1	5–10	5–10	80%	Always	Good: Productivity, quality, repetitive work, terminology.	Good for tech docs and repetition.
eng-nld-oracle-avg	Freelance (FT)	C2	>10	2–5	40%	Always	Bad: less creative, less productive, often wrong	Bad MT is worse than no MT for specialized domains.
eng-nld-oracle-slow	Freelance (FT)	C2	>10	2–5	60%	Often	Good: Productivity, repetitive work. Bad: cultural references.	More productivity at the cost of idioms and cultural factors.
eng-nld-unsup-fast	Freelance (FT)	C2	5–10	2–5	40%	Often	Good: all. Bad: often wrong, worse quality.	PE makes you less in touch with the texts and often poorly paid.
eng-nld-unsup-avg	Freelance (FT)	C2	5–10	2–5	60%	Sometimes	Good: Productivity, quality, repetitive work, terminology. Bad: wrong.	Practical but less effective for longer passages.
eng-nld-unsup-slow	Freelance (FT)	C2	>10	2–5	40%	Always	Good: repetitive work, productivity, terminology	Improves consistency and productivity if applied well.
eng-nld-sup-fast	Freelance (FT)	C2	>10	5–10	60%	Often	Good: repetitive work, creativity, terminology	Useful, but worries about job loss
eng-nld-sup-avg	Freelance (FT)	C2	>10	10	60%	Sometimes	Good: terminology, creativity	Useful for inspiration on better translations
eng-nld-sup-slow	Freelance (FT)	C1	5–10	5–10	80%	Always	Good: repetitive work, productivity	Better productivity at the cost of creativity.

Identifier	Freq. Issues	MT quality	MT fluency	MT accuracy	High. accurate	High. useful	Interface clear	Task difficult	Highlights statements					
									$\uparrow$ Speed?	$\uparrow$ Quality?	$\uparrow$ Effort?	Influence	Spot errors	$\uparrow$ Enjoy?
eng-ita-nohigh-fast	inflection,additions,omissions	4	0.8	0.8	–	–	5	1	–	–	–	–	–	–
eng-ita-nohigh-avg	multiple	3	0.6	0.4	–	–	2	4	–	–	–	–	–	–
eng-ita-nohigh-slow	terminology,omissions	3	0.8	0.8	–	–	1	5	–	–	–	–	–	–
eng-ita-oracle-fast	inflection,terminology	5	0.4	0.8	4	4	4	5	5	2	1	1	1	4
eng-ita-oracle-avg	syntax,terminology,omissions,no context	3	0.4	0.6	2	1	2	3	1	1	4	1	1	1
eng-ita-oracle-slow	syntax,no context	3	0.6	0.6	2	2	2	5	1	1	1	1	4	1
eng-ita-unsup-fast	omissions	3	0.8	0.6	3	2	4	5	3	3	3	2	2	2
eng-ita-unsup-avg	syntax,terminology,no context	3	0.6	0.6	3	3	3	5	2	3	2	1	1	3
eng-ita-unsup-slow	syntax,inflection,terminology,omissions	3	0.4	0.6	2	2	3	4	2	2	3	3	4	4
eng-ita-sup-fast	syntax,terminology,no context	3	0.4	0.4	2	1	2	2	1	1	3	1	2	2
eng-ita-sup-avg	syntax,terminology,no context	3	0.4	0.4	2	2	3	5	3	2	4	3	3	4
eng-ita-sup-slow	syntax,terminology,omissions,no context	3	0.6	0.6	2	2	1	2	2	1	1	4	4	1
eng-nld-nohigh-fast	syntax,terminology,omissions,no context	3	0.2	0.4	–	–	4	4	–	–	–	–	–	–
eng-nld-nohigh-avg	syntax,terminology,omissions,no context	2	0.4	0.6	–	–	4	5	–	–	–	–	–	–
eng-nld-nohigh-slow	terminology,omissions,no context	2	0.2	0.4	–	–	3	5	–	–	–	–	–	–
eng-nld-oracle-fast	syntax,inflection,terminology	3	0.6	0.6	2	1	3	2	2	2	2	1	1	1
eng-nld-oracle-avg	syntax	3	0.8	0.6	4	3	3	4	3	3	3	3	2	3
eng-nld-oracle-slow	syntax,terminology	3	0.6	0.4	3	1	3	4	1	1	1	1	1	3
eng-nld-unsup-fast	terminology,additions,omissions	3	0.6	0.8	3	2	4	4	1	3	1	1	2	1
eng-nld-unsup-avg	multiple	3	0.6	0.6	4	3	2	4	3	3	4	3	2	3
eng-nld-unsup-slow	syntax,terminology,omissions	1	0.4	0.4	2	4	1	4	4	4	3	2	2	3
eng-nld-sup-fast	terminology,omissions,no context	3	0.6	0.4	2	2	3	5	1	1	5	3	1	1
eng-nld-sup-avg	syntax,additions,no context	3	0.4	0.6	2	2	2	4	1	1	1	1	2	3
eng-nld-sup-slow	multiple	5	0.8	1	4	3	2	5	3	3	2	2	2	4

Table 10: **Top:** Sample of pre-task questionnaire results. **Bottom:** Sample of post-task questionnaire results. YoE = years of experience. Post-task statements use a 1–Strongly disagree to 5–Strongly agree scale.

Modalities		English→Italian			English→Dutch			Both		
		Bio	**Social**	**Both**	**Bio**	**Social**	**Both**	**Bio**	**Social**	**Both**
Oracle and	**Sup.**	0.17	0.32	0.25	**0.38**	0.29	0.34	0.26	0.29	0.29
	Unsup.	0.14	0.30	0.20	**0.31**	0.27	0.28	0.22	0.29	0.24
Supervised and	**Oracle**	0.19	**0.31**	0.26	0.30	0.26	0.29	0.24	0.29	0.28
	Unsup.	0.19	**0.33**	0.25	0.28	0.24	0.25	0.24	0.29	0.25
Unsupervised and	**Oracle**	0.22	0.32	0.27	**0.35**	0.30	0.33	0.28	0.31	0.30
	Sup.	0.22	0.37	0.30	**0.39**	0.27	0.33	0.30	0.31	0.32

Table 11: Average highlight agreement proportion between different modalities across language pairs and domains (Section 4.2). Scores are normalized to account for the relative frequency of highlight modalities compared to the mean highlight frequency for the current language and domain combination.

# Doc.-Seg.	Error Type	Has Highlight			% Post-edited			
		Oracle	**Unsup.**	**Sup.**	**No High.**	**Oracle**	**Unsup.**	**Sup.**
1–8	Wrong number	NLD	Both	Both	67	**83**	**83**	**83**
13–6	Remove negation	ITA	Both	Both	**50**	33	33	**50**
16–3	Title literal translation	Both	Both	Both	83	**100**	**100**	**100**
20–1	Wrong acronym	NLD	Both	ITA	0	**33**	**33**	**33**
20–7	Wrong acronym (1)	Neither	Both	Neither	0	**58**	50	25
20–7	Wrong acronym (2)	NLD	Both	ITA	0	**58**	50	25
22–1	Name literal translation	Both	Both	Both	50	50	**83**	67
23–4	Addition	NLD	Both	Neither	**100**	**100**	83	50
31–2	Wrong acronym	NLD	Both	Neither	17	**33**	17	**33**
34–7	Numbers swapped	NLD	Both	NLD	17	50	33	**67**
37–4	Verb polarity inverted	Both	Both	Both	67	**83**	67	**83**
43–5	Wrong name	Both	Both	Both	50	**83**	67	**83**
48–5	Wrong term	NLD	Both	NLD	67	50	**83**	**83**
	Total	65	**100**	62	44	**63**	60	60

Table 12: Highlighting and post-editing statistics for manual critical errors (Section 3.3). Labels in **Has Highlight** columns indicate whether the error was highlighted in Both, only one (ITA or NLD) or Neither directions. Total scores represent the percentage of detected errors (13 errors, 6 editors per highlight modality).

Domain	Modality	$P(H)$	$P(E)$	$P(E\|H)$	$P(E\|\neg H)$	$\Lambda_H(E)$	$P(H\|E)$	$P(H\|\neg E)$	$\Lambda_E(H)$	$F1_H$
English→Italian										
	Random	.12	–	– /.02	– /.02	– /1.0	– /.11	– /.13	– /0.8	– /.03
	No High.	–	.02	–	–	–	–	–	–	–
Biomed.	**Oracle**	.08	.07	.26/.08	.05/.02	5.2/4.0	.30/.26	.06/.08	5.0/3.2	.28/.12
	Unsup.	.16	.10	.18/.06	.08/.02	2.2/3.0	.29/.36	.14/.15	2.0/2.4	.22/.10
	Sup.	.11	.12	.18/.05	.11/.02	1.6/2.5	.16/.23	.10/.10	1.6/2.3	.17/.08
	Random	.20	–	– /.09	– /.09	– /1.0	– /.21	– /.20	– /1.0	– /.13
	No High.	–	.09	–	–	–	–	–	–	–
Social	**Oracle**	.25	.20	.42/.23	.13/.04	3.2/5.7	.52/.66	.18/.21	2.8/3.1	.46/.34
	Unsup.	.17	.18	.35/.19	.14/.07	2.5/2.7	.33/.37	.14/.15	2.3/2.4	.34/.25
	Sup.	.15	.21	.38/.23	.18/.06	2.1/3.8	.27/.39	.11/.12	2.4/3.2	.32/.29
English→Dutch										
	Random	.17	–	– /.12	– /.10	– /1.2	– /.19	– /.17	– /1.1	– /.15
	No High.	–	.10	–	–	–	–	–	–	–
Biomed.	**Oracle**	.21	.08	.21/.20	.05/.08	4.2/2.5	.52/.41	.18/.18	2.8/2.2	.30/.27
	Unsup.	.23	.09	.17/.17	.07/.08	2.4/2.1	.43/.38	.21/.21	2.0/1.8	.24/.23
	Sup.	.12	.08	.20/.21	.06/.09	3.3/2.3	.30/.25	.11/.11	2.7/2.2	.24/.23
	Random	.16	–	– /.22	– /.19	– /1.1	– /.19	– /.16	– /1.1	– /.17
	No High.	–	.19	–	–	–	–	–	–	–
Social	**Oracle**	.19	.12	.33/.39	.07/.15	4.7/2.6	.54/.39	.15/.15	3.6/2.6	.41/.39
	Unsup.	.15	.13	.25/.33	.11/.17	2.2/1.9	.30/.26	.13/.12	2.3/2.1	.27/.29
	Sup.	.12	.10	.30/.36	.08/.17	3.7/2.1	.36/.23	.10/.10	3.6/2.3	.33/.28

Table 13: Highlighting (H) and editing (E) statistics for each domain, modality and translation direction combination ($n = 3$ post-editors per combination). Values after slashes are adjusted by projecting highlights of the specified modality over edits from **No Highlight** translators to estimate highlight-induced editing biases (Section 4.2). A **Random** baseline is added by projecting random highlights matching the average frequency over all modalities for specific domain and translation direction settings.

Domain	Speed	$P(H)$	$P(E)$	$P(E\|H)$	$P(E\|\neg H)$	$\Lambda_H(E)$	$P(H\|E)$	$P(H\|\neg E)$	$\Lambda_E(H)$	$F1_H$
English→Italian										
	Fast		.04/.01	.12/.02	.03/.01	4.0/2.0	.30/.27	.08/.11	3.7/2.4	.17/.04
Biomed.	Avg.	.09	.10/.05	.27/.12	.09/.04	3.0/3.0	.22/.30	.07/.11	3.1/2.7	.24/.17
	Slow		.09/.02	.21/.04	.08/.01	2.6/4.0	.19/.26	.07/.11	2.7/2.3	.20/.07
	Fast		.11/.07	.30/.20	.07/.04	4.2/5.0	.40/.52	.11/.16	3.6/3.2	.34/.29
Social	Avg.	.14	.23/.14	.48/.32	.18/.10	2.6/3.2	.30/.42	.09/.15	3.3/2.8	.37/.36
	Slow		.17/.05	.39/.14	.14/.03	2.7/4.6	.31/.54	.11/.17	2.8/3.1	.35/.22
English→Dutch										
	Fast		.03/.02	.11/.05	.02/.01	5.5/5.0	.48/.61	.13/.18	3.6/3.3	.18/.09
Biomed.	Avg.	.14	.11/.19	.20/.30	.10/.17	2.0/1.7	.25/.29	.13/.16	1.9/1.8	.22/.29
	Slow		.12/.10	.26/.23	.10/.07	2.6/3.2	.29/.42	.12/.16	2.4/2.6	.27/.30
	Fast		.06/.07	.19/.21	.04/.04	4.7/5.2	.37/.47	.10/.13	3.7/3.6	.25/.29
Social	Avg.	.12	.17/.32	.32/.48	.15/.29	2.1/1.6	.22/.23	.10/.12	2.2/1.9	.26/.31
	Slow		.18/.18	.38/.40	.15/.14	2.5/2.8	.25/.34	.09/.11	2.7/3.0	.30/.37

Table 14: Highlighting (H) and editing (E) statistics for each domain, and translation direction across translator speeds ($n = 4$ post-editors per combination, regardless of highlight modality). Values after slashes are adjusted by projecting highlights of the specified modality over edits from **No Highlight** translators to estimate highlight-induced editing biases (Section 4.2).

Language	MQM Category	MT		No Highlight		Oracle		Unsupervised		Supervised	
		Maj.	Min.	Maj.	Min.	Maj.	Min.	Maj.	Min.	Maj.	Min.
Italian	Accuracy - Addition	0	1	0	0	0	0	0	0	1	1
	Accuracy - Mistranslation	21	22	10	12	4	8	24	17	17	17
	Accuracy - Inconsistency	2	4	1	3	2	2	1	3	0	2
	Accuracy - Omission	2	0	0	0	0	1	4	1	1	2
	Accuracy - Untranslated	1	4	1	2	0	1	1	1	3	2
	Style - Inconsistent Style	0	0	0	0	0	0	0	0	0	0
	Style - Readability	17	25	5	30	0	12	4	34	1	29
	Style - Wrong Register	0	8	0	3	0	3	1	1	3	2
	Linguistic - Grammar	6	15	2	16	0	5	3	12	2	12
	Linguistic - Punctuation	1	13	0	9	0	3	1	6	0	3
	Linguistic - Spelling	5	3	0	4	0	3	3	2	0	1
	Total	55	95	19	79	6	38	42	77	28	71
Dutch	Accuracy - Addition	0	1	0	2	0	3	0	2	0	1
	Accuracy - Mistranslation	25	34	18	25	23	27	12	31	16	29
	Accuracy - Inconsistency	0	0	0	2	0	2	0	2	0	5
	Accuracy - Omission	3	1	1	1	2	1	1	1	4	2
	Accuracy - Untranslated	4	4	1	1	1	4	1	3	0	2
	Style - Inconsistent Style	2	0	0	5	1	7	0	2	0	9
	Style - Readability	1	27	1	20	0	13	2	15	6	41
	Style - Wrong Register	0	2	0	3	0	3	0	1	1	0
	Linguistic - Grammar	3	19	2	14	3	23	2	6	3	12
	Linguistic - Punctuation	0	6	0	3	0	4	0	2	0	3
	Linguistic - Spelling	1	1	1	1	2	1	0	1	0	0
	Total	39	95	24	77	32	88	18	66	30	104

Table 15: MQM error counts averaged across $n = 3$ translators per highlight modality for every translation direction. A description of MQM categories is available in Figure 5.

Method	DivEMT				QE4PE			
	En→It		En→Nl		En→It		En→Nl	
	AP	AU	AP	AU	AP	AU	AP	AU
LOGPROBS (Fomicheva et al., 2020)	0.18	0.18	0.19	0.19	0.10	0.09	0.09	0.09
LOGPROBS$_{\text{MCD VAR}}$ (Fomicheva et al., 2020, **Unsup.**)	0.41	0.41	0.42	0.42	0.23	0.23	0.31	0.31
XCOMET-XXL (Guerreiro et al., 2024, **Sup.**)					0.16	0.23	0.19	0.28
AVG. **Oracle** SINGLE TRANSLATOR	–	–	–	–	0.53	0.73	0.55	0.75

Table 16: Average Precision (AP) and Area Under the Precision-Recall Curve (AU) between metrics and error spans derived from human post-editing. We use mBART 1-to-50 (Tang et al., 2021) and NLLB 3B (NLLB Team et al., 2024) respectively for DivEMT and QE4PE. For DivEMT, a single post-editor is available for computing the agreement, while for QE4PE we use consensus-based **Oracle** highlights. For QE4PE, we report the average agreement between individual oracle post-editors and their consensus as an agreement upper bound.

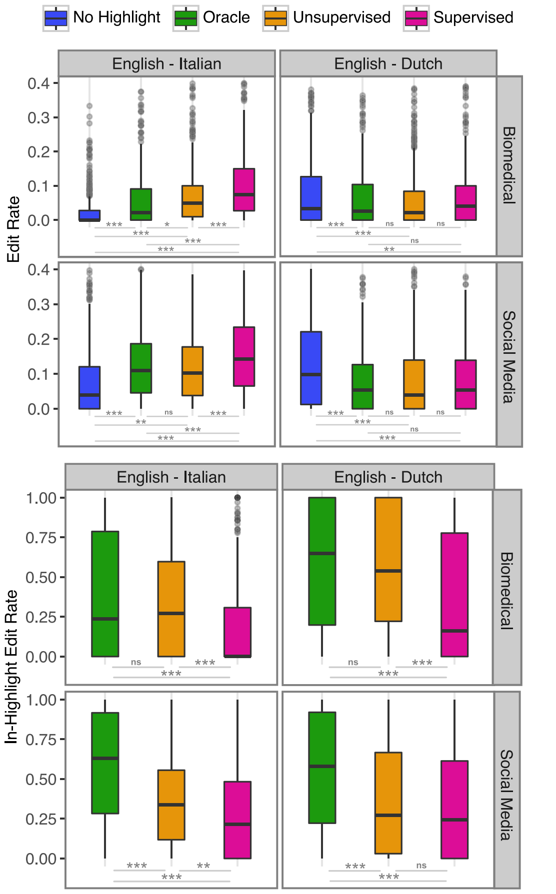


Figure 6: **Top:** Post-editing rate across highlight modalities, domains and directions. **Bottom:** Proportion of edits in highlighted spans across highlight modalities. $*** = p < 0.001$, $** = p < 0.01$, $* = p < 0.05$, ns = not significant with Bonferroni correction.

Do post-edits in QE: / *Also happen in:*

	en→it Oracle	en→it Sup'd	en→it Unsup'd	en→nl Oracle	en→nl Sup'd	en→nl Unsup'd
Oracle	79%	75%	66%	72%	75%	67%
Supervised	82%	82%	73%	78%	82%	72%
Unsupervised	76%	76%	71%	75%	78%	70%
No highlight	60%	62%	52%	79%	82%	75%

Figure 7: Post-editing agreement across various modalities (Section 4.2). Results are averaged across all translator pairs for the two modalities ($n = 3$ intra-modality, $n = 9$ inter-modality for every language) and all segments.

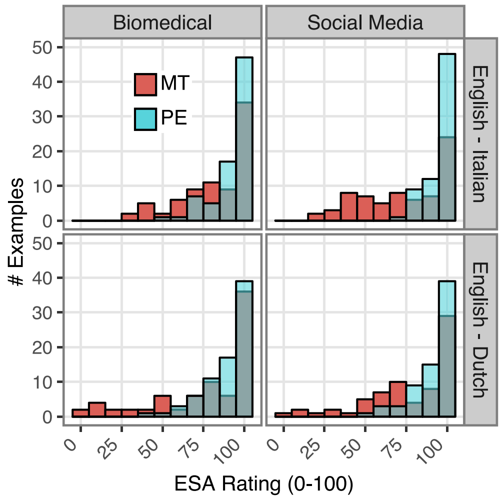


Figure 8: ESA ratings for MT outputs and post-edits across domains and translation directions.

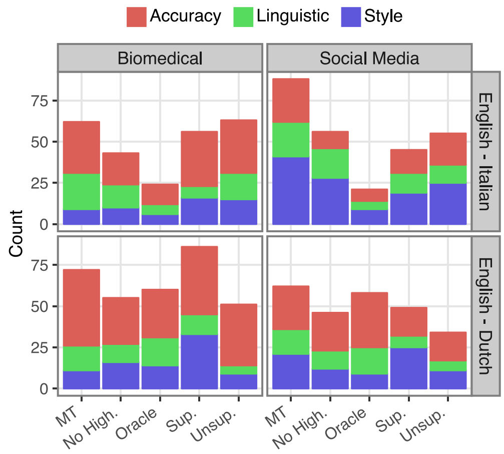


Figure 9: Distribution of MQM error categories for MT and post-edits across highlight modalities for the two translation directions and domains of QE4PE.

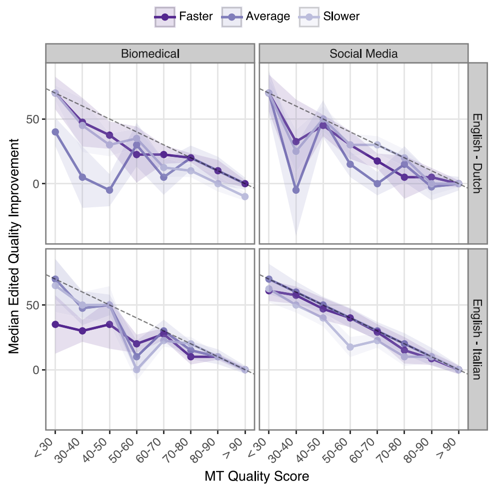


Figure 10: Median ESA quality improvement following post-editing for segments at various initial MT quality levels across translators' speed groups, showing no clear quality trends across editors' productivity levels.

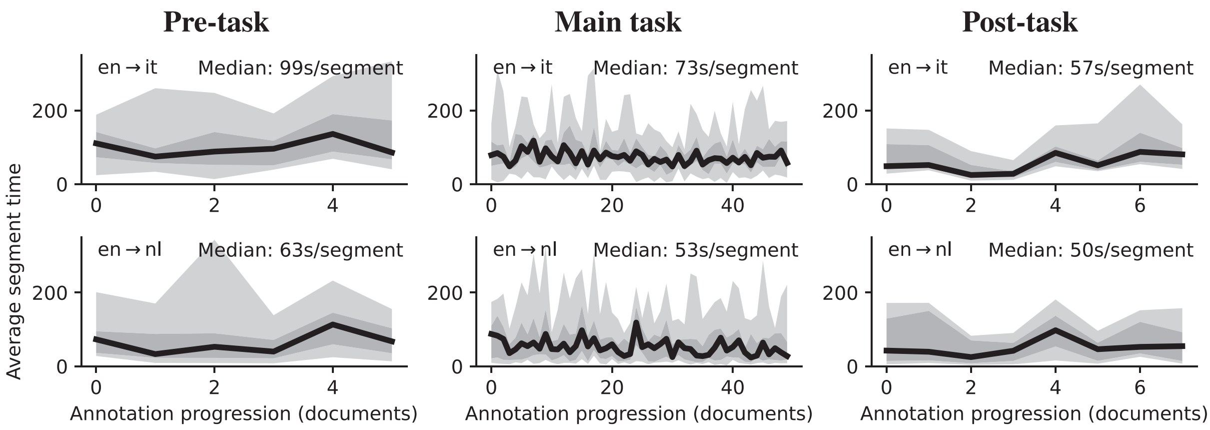


Figure 11: Segment-level post-editing time with respect to post-editor progression. Values are medians across all annotators. Light gray area is min-max values, dark gray represents 25%–75% quantiles. The annotators do not became considerably faster with the task progression, likely due to the simplicity of the task and the high post-editing proficiency of professional post-editors. The high variability in editing times motivates the careful group assignments performed using PRE task edit logs.

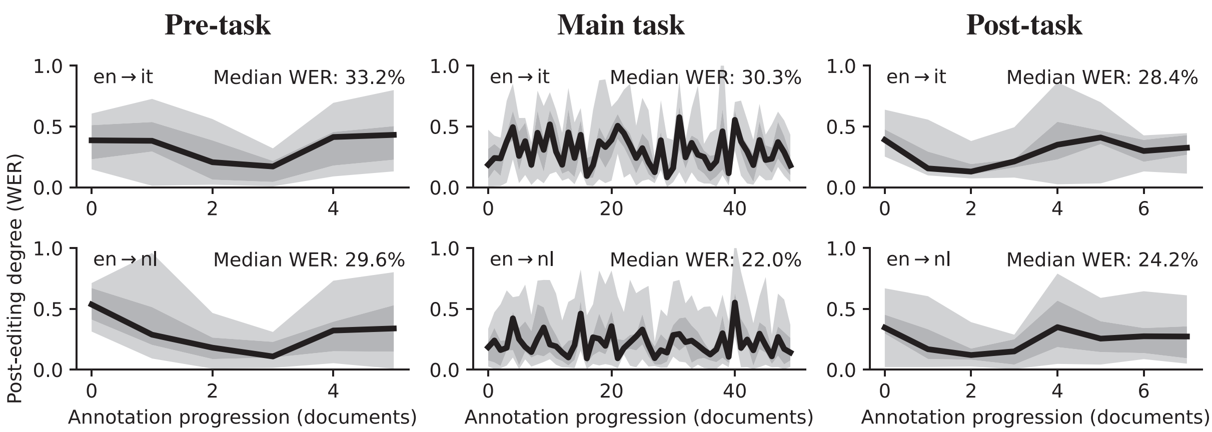


Figure 12: Editing proportion, measured by word error rate between MT and post-edited texts, with respect to post-editor progression. Values are medians across all post-editors.