International Journal of Linguistics, Literature and Translation
ISSN: 2617-0299 (Online); ISSN: 2708-0099 (Print)
DOI: 10.32996/ijllt
Journal Homepage: www.al-kindipublisher.com/index.php/ijllt

IJLLT
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Cognitive Effort and Perceptions in Translation Revision and Post-Editing

Majdi Alzahrani

Assistant Professor, King Abdulaziz University, College of Science and Arts, Rabigh, Saudi Arabia
Corresponding Author: Majdi Alzahrani, **E-mail**: mmalzahrani3@kau.edu.sa

| ABSTRACT

This study investigates the correlation between cognitive effort and translator perceptions during human translation (HT) revision and machine translation (MT) post-editing. Seven participants revised and post-edited two texts without knowing whether the texts had been produced by a human or a machine. Eye-tracking and keylogging captured fixation and pause durations (FD and PD) to measure cognitive effort, while a one-question survey assessed participants' perceptions of translation type. Linear Mixed Models (LMM) were used to analyze fixation and pause duration data at both individual and group levels. FD results showed that MT revisions generally required greater cognitive effort for source texts but slightly less for target texts, while PD showed no uniform trend toward either lower or higher cognitive effort. Survey responses indicated a preference for identifying HT with higher confidence, while MT was frequently misclassified as HT. This suggests a possible correlation between misclassification of MT as HT and higher fixation durations, indicating that perception might influence cognitive processing. The study offers insights into translation practice, particularly by highlighting how assumptions about whether a text was produced by HT or MT may affect revision and post-editing decisions.

| KEYWORDS

Cognitive Effort, Perception, Human Translation, Machine Translation, Revision, Post-Editing

| ARTICLE INFORMATION

ACCEPTED: 01 May 2025 **PUBLISHED:** 02 June 2025 **DOI:** 10.32996/ijllt.2025.8.6.1

1. Introduction

Translation practices have evolved with advancements in machine translation (MT) technologies (Doherty, 2016). MT, along with other supporting technologies such as computer-aided translation (CAT) tools, is now considered an integral part of professional translation workflows and environments. This has transformed translators' roles from traditional mediators between the source text and target audience (that is, translating from scratch with barely any technological involvement) to revisers of human translation (HT) or post-editors of MT output—two roles that require different skill sets and cognitive demands (see, e.g., Pym, 2013). Additionally, recent developments in translation technology, such as translation memory (TM) and, arguably most revolutionary, artificial intelligence (AI), have accelerated engagement in revision and post-editing (Robert et al., 2022), introducing diverse cognitive and practical demands. Furthermore, as MT quality continues to improve, at least in some language pairs, the boundaries between HT and MT have become less distinct in appearance, particularly in terms of adherence to linguistic norms. This resemblance can conceal translation errors, which, as a result, may require more time and greater cognitive effort to post-edit (Daems et al., 2017). In other words, such closeness in form and meaning between HT and MT does not necessarily imply equal quality or reliability; therefore, post-editing remains essential to ensure translation quality (see, e.g., Jakobsen, 2018).

These evolving roles and practices in translation call for a closer examination of how revision and post-editing differ in process and cognitive demand. Revision and post-editing serve to ensure the accuracy, coherence, and fidelity of translated texts; however, they differ in approach, guidelines, and cognitive demands. Revision typically involves refining human-produced translations with an emphasis on linguistic quality and creativity, while post-editing focuses on correcting machine-generated texts, often minimizing stylistic intervention (Daems et al., 2020; Nitzke & Gros, 2020). According to Daems and Macken (2020), omissions and additions occur more frequently in revision, whereas repeatable and predictable errors occur more frequently in post-editing. Additionally, the two tasks differ in important respects: revision emphasizes critical reading and feedback skills, while post-editing demands familiarity with MT systems and common MT errors (Robert et al., 2022).

While post-editing is generally perceived as faster—mainly due to the predictability of MT errors—this advantage depends heavily on factors such as output quality, especially given that MT performance varies across language pairs. In effect, low-quality translations may require extensive rewriting, increasing cognitive effort and reducing productivity (Koponen, 2016). Furthermore, perceptions about whether a text was produced by a human or a machine have been shown to influence editing behavior and translation quality (Briva-Iglesias & O'Brien, 2024). While some studies have shown that assumptions about translation production can affect the editing process (Daems & Macken, 2020), the direct relationship between perceptions and cognitive effort has not been fully addressed in the literature. Therefore, this study aims to examine whether cognitive effort and perceptions correlate during revision and post-editing. It uses eye-tracking and keylogging to measure cognitive effort objectively and employs a post-task survey to gather subjective perceptions.

2. The Experiment

This section outlines the experimental setup used to investigate the correlation between cognitive effort and perception during the revision and post-editing of translated texts. It provides an overview of the number of participants and their demographic characteristics, describes the instruments used to capture data, the metrics applied to measure cognitive effort, and the materials selected for the study. It also outlines the three-part experimental procedure, which included a training session, a revision/post-editing task, and a post-task survey.

2.1 Participants

Seven undergraduate students specializing in translation were invited to participate in the experiment. All participants shared the same linguistic background, with Arabic (L1) as their native and target language, and English (L2) as their second and source language. They were right-handed males, aged between 20 and 22 years (M = 21.14; SD = 0.64), with no history of neurological or psychological disorders, and were not taking any medication. Two participants had normal vision, and five had corrected-to-normal vision. A detailed briefing on the experiment's purpose and procedures was provided to each participant, including assurances of data confidentiality, privacy protection, and their right to withdraw at any time without penalty. All participants were given a consent form and voluntarily signed it.

2.2 Instruments

Two instruments were employed to collect data: the Gazepoint GP3 HD Desktop Eye-Tracker and the Translog-II keylogging software. The Gazepoint Eye-Tracker operates at a frequency range of 60–150 Hz and delivers a viewing angle resolution between 0.5° and 1°, allowing it to capture eye movements at an interval of 6.7 milliseconds. Its dedicated software, Gazepoint Analysis UX Edition, enables the assignment of Areas of Interest (AOIs) and supports various analytical measurements, including the extraction of fixation and gaze data into .CSV files.

The Translog-II software, developed specifically for Translation Process Research (TPR), records keyboard activities during the post-editing task. It consists of two components: Translog-II Supervisor, which researchers use to set up and manage experiments, and Translog-II User, which records keyboard and mouse activity throughout the task and generates time-stamped log files that can be extracted and analyzed. These log files contain time-stamped keylogging data, allowing researchers to extract and analyze pause periods for further insights into the translation process (Carl, 2012).

2.3 Metrics

The metric data extracted from the eye-tracker and keylogging tools were fixation duration (FD) and pause duration (PD), respectively. In Translation Process Research (TPR), FD refers to the total time a word or area of interest is fixated on during translation (Rydning & Lachaud, 2010). Prior research has shown that FD correlates with cognitive effort (Saldanha & O'Brien, 2014), and that a 200-millisecond is commonly adopted in TPR studies (Alves et al., 2009).

Pause duration (PD) refers to the length of pauses that occur during the translation task, i.e., the more difficult the task, the more cognitive effort (and pauses) are required (Koponen, 2016). PD is also commonly employed to explore translators' cognitive effort in translation and related tasks such as revision and post-editing. A 300-millisecond threshold was applied, which falls within the range typically reported in similar studies (Jakobsen, 2017).

2.4 Materials

The experiment involved two English source texts and their respective Arabic translations. The first source text (ST1) is an excerpt from a news report published by The Associated Press (2023) and was translated by the researcher (i.e., human translation). The second source text (ST2), on the other hand, was written by the researcher to closely match ST1 in terms of style and structure, and was translated using a machine translation system (Google Translate).

Table 1. Source and Target Texts Used in the Experiment

ST1	ST2
A cargo jet headed to Belgium from New York had to turn around mid-flight after a horse escaped its stall and got loose in the hold, according to air traffic control audio. The Boeing 747 operated by Air Atlanta Icelandic had just started its flight across the Atlantic Ocean on Nov. 9 when the pilot radioed air traffic control in Boston and said that a horse on board had escaped its stall.	An oil tanker sailing from Houston to Rotterdam encountered an unexpected issue mid-voyage when an engine failure disrupted its progress, as reported in maritime radio communications. The massive vessel, operated by Global Oil Shipping, was traversing the Atlantic Ocean on Nov. 8 when the ship's captain contacted maritime control in Amsterdam, alerting them to the engine room failure, which necessitated immediate attention and potential repairs before continuing the journey.
HT	**MT**
طائرة نقل متجه إلى بلجيكا من نيويورك اضطرت إلى العودة عقب وصولها نصف مسافة السفر بعد أن فر حصاناً من المربط وأصبح دون قيد، وفقاً لحركة تحكم الملاحة الجوية. طائرة بوينق ٧٤٧ يتم تشغيلها بواسطة أتلانتيك الأيسلندية الجوية كانت قد بدأت رحلتها عبر المحيط الأطلسي يوم ٩ نوفمبر عندما تواصل طائرها بحركة الملاحة الجوية في بوسطن وأخبرهم أن حصانا على متن الرحلة كان قد فر من المربط الجوي.	واجهت ناقلة نفط تبحر من هيوستن إلى روتردام مشكلة غير متوقعة في منتصف الرحلة عندما أدى عطل في المحرك إلى تعطيل تقدمها، كما ورد في الاتصالات اللاسلكية البحرية. وكانت السفينة الضخمة، التي تديرها شركة Global Oil Shipping، تعبر المحيط الأطلسي في 8 نوفمبر عندما اتصل قبطان السفينة بالمراقبة البحرية في أمستردام، لتنبيههم إلى عطل في غرفة المحرك، مما استلزم اهتمامًا فوريًا وإصلاحات محتملة قبل مواصلة الرحلة.

2.5 Procedure

The experiment consisted of three parts: an instruction and training session, the revision/post-editing task, and the post-task survey. During the first part, participants were briefed on the study's objectives and procedures, and were asked to complete a demographic form and sign an informed consent document. Furthermore, a 25-minute training session was conducted to guide participants through the eye-tracker calibration process and explain the proper seating position (65 cm from the screen) required for accurate data capture, according to the GP3 eye-tracker manual (2025). Next, participants performed eye-tracker calibration and then completed the assigned revision and post-editing tasks, while both Gazepoint Analysis UX Edition and Translog-II software recorded the process. The Translog-II main interface was divided into two vertically stacked windows: one displayed the source text, and the other showed the translation to be revised or post-edited. After each task, participants were given a one-question survey to assess their perceptions of the translation type.

3. Processing of Data

This section explains how the collected data were prepared for analysis. It begins with how fixation data were assessed for quality and compiled into datasets for statistical comparison. Then, it outlines how pause data were extracted and arranged for statistical testing. Finally, it explains how participants' survey responses were used as a subjective measure of perceived translation type.

3.1 Fixation Data

After the experiment was concluded, an initial review of each recording was conducted to confirm the presence of saccades, i.e., to ensure that fixation data had been captured for both revision and post-editing tasks. Next, using the Gazepoint software, two Areas of Interest (AOIs) were designated for each participant's recording: source text (ST) and target text (TT). Fixation durations were then extracted via the software, including fixations on the AOIs. As a result, three Excel files were generated, labeled as follows: Data Summary Report (DSR), Fixation Samples (FS), and Gaze Samples (GS). The DSR provided summarized eye-tracking data such as total fixation durations and counts, but without disclosing the threshold at which these data were calculated. The FS contained recorded fixation samples only within the designated AOIs. The GS contained all fixation samples, whether short or long, including both on-screen and off-screen fixations.

Before analysis, an eye-tracking data quality measure was applied: the calculation of mean fixation duration (MFD). As a quality metric, MFD indicates whether the fixation data are reliable enough to yield meaningful results (Hvelplund, 2014). It was calculated by dividing the total fixation duration for each AOI by its fixation count. In line with prior studies using MFD, results below 200 milliseconds are considered unacceptable and should therefore be excluded from further analysis (see, e.g., Pavlović & Jensen, 2009; Hvelplund, 2011). However, as shown in Table 2, all MFD values were found to be above this threshold; hence, the fixation data were considered of acceptable quality for each participant.

Table 2. Mean Fixation Duration (milliseconds).

	Human		Machine	
Participant	ST	TT	ST	TT
1	0.298	0.414	0.313	0.308
2	0.291	0.346	0.298	0.362
3	0.366	0.332	0.372	0.351
4	0.318	0.368	0.35	0.346
5	0.345	0.358	0.455	0.358
6	0.339	0.387	0.326	0.366
7	0.353	0.319	0.410	0.360

After extracting, processing, and confirming the quality of the fixation data, four fixation duration (FD) datasets were compiled for each participant: FD/ST and FD/TT for HT, and FD/ST and FD/TT for MT. This structure enabled the application of a statistical model—namely, Linear Mixed Models (LMM)—to compare HT revision with MT post-editing for each participant and to determine whether cognitive effort varies between the two tasks. For example, this involved comparing participant one's (P1) FD/ST data for HT against FD/ST data for MT, or FD/TT for HT against FD/TT for MT. For the purpose of this study, this type of analysis is referred to as individual-level analysis. Furthermore, to enable analysis of FD at the group level, each type of FD (e.g., FD/ST for HT) was compiled into a single file containing data from all participants and organized into datasets for comparison with its counterpart (i.e., [GFD/ST for HT vs. GFD/ST for MT]; [GFD/TT for HT vs. GFD/TT for MT]). To visualize the group-level data, Figures 1 and 2 display GFD/ST and GFD/TT, respectively, over task time for both HT and MT.

3.2 Pause Data

Extracted as XML files from Translog-II, the keylogging data included timestamps and event types for each participant. From this, pause durations of 300 milliseconds or more were compiled into two PD/TT datasets: one for HT and one for MT. This structure enabled statistical analysis at the individual level to compare MT and HT data for each participant and determine whether cognitive effort differences exist between the two tasks.

Furthermore, to enable group-level statistical analysis, each PD type was combined into a single file containing data from all participants and organized for comparison with its counterpart (e.g., [GPD/TT for HT vs. GPD/TT for MT]). To visualize the group-level data, Figure 3 displays GPD/TT over task time for both HT and MT.

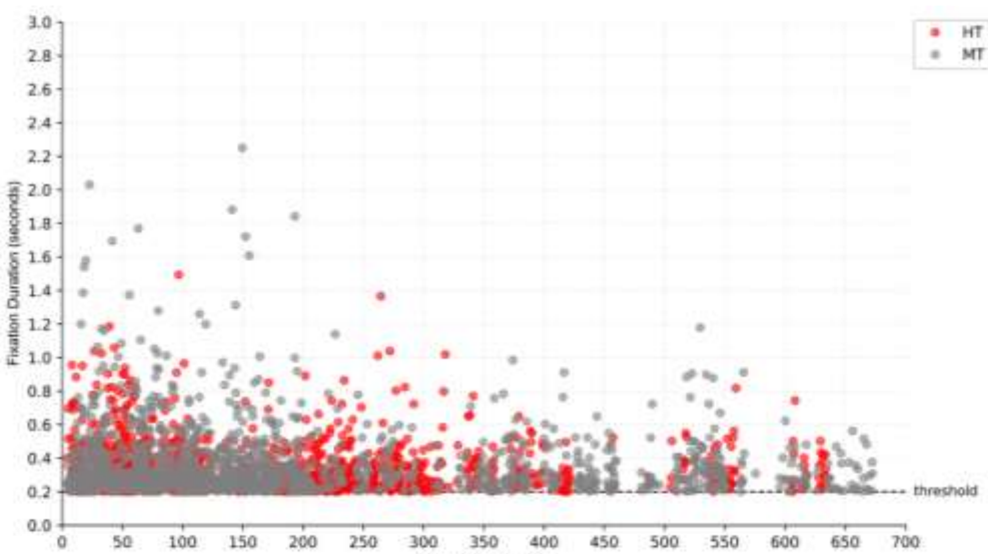

Figure 1. Source Text Fixation Duration Across Task Time for Human Translation (HT) and Machine Translation (MT).

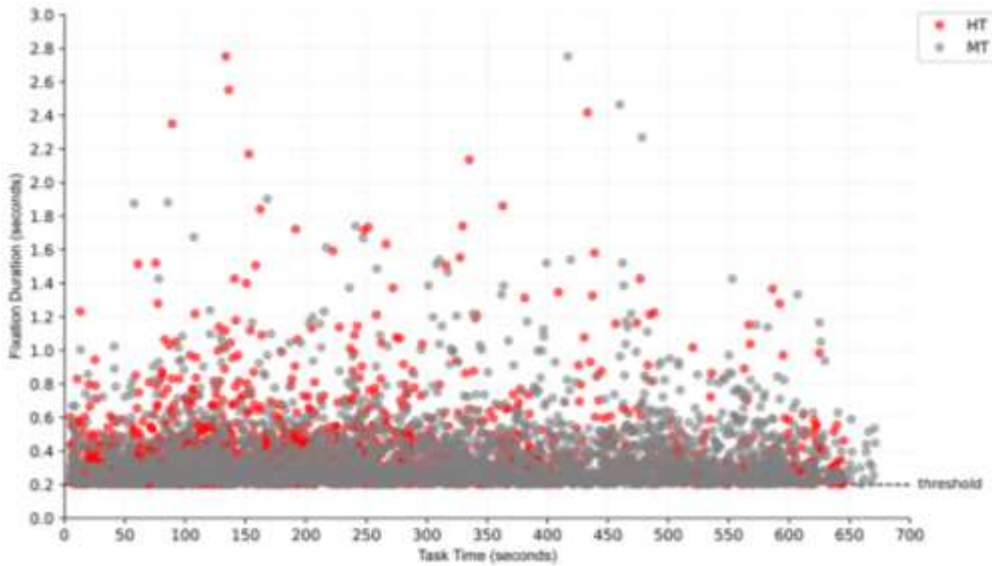

Figure 2. Target Text Fixation Duration Across Task Time for Human Translation (HT) and Machine Translation (MT).

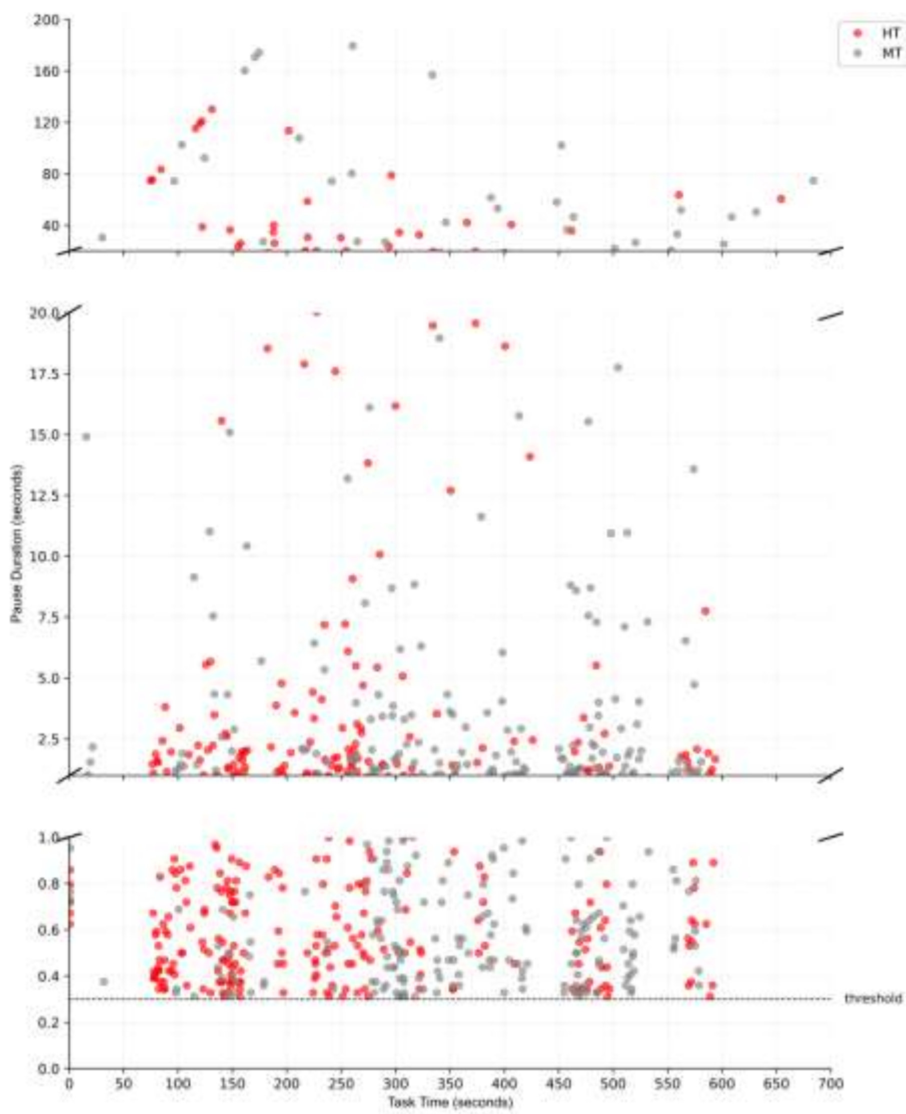

Figure 3. Pause Duration Across Task Time for Human Translation (HT) and Machine Translation (MT).

3.3 Subjective Measure

A single-question survey was administered to participants following each revision or post-editing task to capture perceptions of the translation type. Participants were asked: "Do you think the target text was created by a human or generated by machine translation?" The answer options were presented on a 5-point scale: 1 = "HT", 2 = "Likely HT", 3 = "Not Sure", 4 = "Likely MT", and 5 = "MT". Responses were collected to compare participants' perceptions with cognitive effort metrics recorded through eye-tracking and keylogging.

4. Findings

This section presents the main findings of the study. It begins with individual-level results for fixation and pause durations, comparing cognitive effort during the revision of HT and the post-editing of MT (see Table 3). It then reports group-level results to identify trends across participants (see Table 4). Finally, it summarizes the survey responses to examine how participants perceived the translation type.

4.1 Fixation Duration Individual-Level

For P1, P2, and P3, there were no statistically significant differences in FD/ST between MT and HT ($p = 0.298$, $p = 0.492$, and $p = 0.734$, respectively). For these participants, FD/ST for MT was marginally higher than for HT by 0.014 seconds (P1), 0.007 seconds (P2), and 0.006 seconds (P3). Thus, there is no evidence of a meaningful difference in cognitive effort between MT and HT for these cases.

In contrast, statistically significant differences in FD/ST were observed for P4, P5, and P7 ($p = 0.020$, $p < 0.001$, and $p = 0.030$, respectively). For P4, FD/ST for MT was 0.033 seconds higher than for HT; for P5, the difference was substantially larger at 0.111 seconds; and for P7, the difference was 0.057 seconds. These results suggest that these participants experienced noticeably higher cognitive effort when revising MT output compared to HT. For P6, FD/ST for MT was 0.012 seconds lower than for HT; however, this difference was not statistically significant ($p = 0.295$), again suggesting no meaningful difference in cognitive effort.

Regarding FD/TT, P1 showed a significant difference, with MT 0.106 seconds lower than HT ($p = 0.001$), suggesting reduced cognitive effort during post-editing compared to revision. For P2 and P3, FD/TT for MT was slightly higher than HT by 0.016 and 0.019 seconds, respectively, but these differences were not statistically significant ($p = 0.121$ and $p = 0.126$).

For P4, FD/TT for MT was 0.022 seconds lower than HT ($p = 0.080$), while for P5, the difference was negligible: 0.001 seconds lower ($p = 0.979$). P6 showed a 0.022-second decrease for MT compared to HT ($p = 0.098$), and P7 showed a 0.040-second increase for MT ($p = 0.085$); however, none of these differences were statistically significant. These results suggest insufficient evidence to conclude a meaningful difference in cognitive effort based on FD/TT for these participants.

In summary, the individual-level analysis of FD indicates inconsistent patterns in cognitive effort between MT and HT tasks, with statistically significant variation observed in only a few cases.

4.2 Pause Duration Individual-Level

Pause duration data for P1 and P5 were excluded from analysis due to inactivity or anomalously long pauses that did not reflect actual task engagement. For the remaining participants, no statistically significant differences in PD/TT between MT and HT were found. For P2, the PD/TT for MT was 0.082 seconds higher than HT ($p = 0.981$); for P3, it was 3.318 seconds higher ($p = 0.437$); for P4, 0.688 seconds higher ($p = 0.880$); for P6, 0.528 seconds higher ($p = 0.749$); and for P7, 0.651 seconds higher ($p = 0.955$). These results indicate no meaningful difference in cognitive effort between MT and HT for these participants based on PD/TT.

4.3 Fixation and Pause Durations Group-Level

The GFD/ST for MT was, on average, 0.025 seconds higher than that for HT, a statistically significant difference ($p < 0.001$). This indicates that revising MT required more cognitive effort than HT, as reflected in the increased fixation durations. In contrast, GFD/TT for MT was 0.012 seconds lower than that for HT, which is also a statistically significant difference ($p = 0.044$). Although the reduction was small, this indicates slightly lower cognitive effort for MT compared to HT in this context. The GPD/TT for MT was, on average, 1.05 seconds longer than that of HT; however, this difference was not statistically significant ($p = 0.479$). This suggests that while MT may sometimes require more cognitive effort, pause duration differences compared to HT are not consistent across participants or tasks.

Table 3. Individual-Level LMM Results: Differences Between HT and MT

Participant	Variable	Coef	SE	*t*	*p*	LCL	UCL
	Intercept	0.299	0.011	28.07	0.000	0.278	0.320
	FD/ST	0.014	0.014	1.043	0.298	-0.013	0.041
1	Intercept	0.414	0.020	20.45	0.000	0.375	0.454
	FD/TT	-0.106	0.031	-3.455	0.001	-0.166	-0.046
	Intercept	***	***	***	***	***	***
	PD/TT	***	***	***	***	***	***
	Intercept	0.291	0.008	37.44	0.000	0.276	0.307
	FD/ST	0.007	0.011	0.688	0.492	-0.014	0.029
2	Intercept	0.346	0.008	43.53	0.000	0.331	0.362
	FD/TT	0.016	0.010	1.552	0.121	-0.004	0.036
	Intercept	7.640	2.652	2.881	0.005	2.391	12.88
	PD/TT	0.082	3.433	0.024	0.981	-6.711	6.875
	Intercept	0.367	0.013	27.28	0.000	0.340	0.393
	FD/ST	0.006	0.017	0.339	0.734	-0.028	0.040
3	Intercept	0.333	0.010	32.16	0.000	0.313	0.353
	FD/TT	0.019	0.012	1.530	0.126	-0.005	0.044
	Intercept	5.380	3.270	1.645	0.102	-1.090	11.85
	PD/TT	3.318	4.254	0.780	0.437	-5.098	11.73
	Intercept	0.318	0.012	27.58	0.000	0.296	0.341
	FD/ST	0.033	0.014	2.334	0.020	0.005	0.060
4	Intercept	0.369	0.010	35.92	0.000	0.349	0.389
	FD/TT	-0.022	0.013	-1.752	0.080	-0.047	0.003
	Intercept	6.029	3.689	1.634	0.104	-1.265	13.32
	PD/TT	0.688	4.551	0.151	0.880	-8.311	9.686
	Intercept	0.345	0.020	17.04	0.000	0.305	0.385
	FD/ST	0.111	0.028	3.883	0.000	0.055	0.167
5	Intercept	0.359	0.018	19.96	0.000	0.323	0.394
	FD/TT	-0.001	0.023	-0.026	0.979	-0.046	0.045
	Intercept	***	***	***	***	***	***
	PD/TT	***	***	***	***	***	***
	Intercept	0.339	0.007	50.29	0.000	0.326	0.352
	FD/ST	-0.012	0.012	-1.048	0.295	-0.036	0.011
6	Intercept	0.388	0.009	42.07	0.000	0.370	0.406
	FD/TT	-0.022	0.013	-1.656	0.098	-0.048	0.004
	Intercept	4.306	1.060	4.062	0.000	2.218	6.394
	PD/TT	0.528	1.649	0.320	0.749	-2.719	3.775
	Intercept	0.353	0.020	17.31	0.000	0.313	0.393
	FD/ST	0.057	0.026	2.177	0.030	0.006	0.109
	Intercept	0.320	0.019	16.88	0.000	0.282	0.357
7	FD/TT	0.040	0.023	1.724	0.085	-0.006	0.086
	Intercept	4.306	1.060	4.062	0.000	2.218	6.394
	PD/TT	0.528	1.649	0.320	0.749	-2.719	3.775

4.4 Perception of Translation Type

For MT, the distribution of participant responses across categories revealed equal preferences for "HT" and "Likely HT," each accounting for 28.57% of the total responses. The remaining categories, "Not Sure," "Likely MT," and "MT", each represented 14.29%, indicating a more balanced uncertainty and lower confidence in identifying MT as the translation type. This distribution highlights variability in participant perceptions and suggests mixed levels of certainty across the categories.

For HT, the analysis of participant responses showed that HT received the highest proportion of responses at 42.86%, reflecting a strong tendency among participants to identify the text as HT. Both "Likely HT" and "Not Sure" accounted for 28.57% each, indicating some level of uncertainty in their judgments. Notably, "Likely MT" and "MT" received no responses, suggesting that participants did not associate the text with MT or express confidence in MT classification. These findings highlight a clear preference for HT identification, with some divided perceptions but no inclination toward MT.

The results of the translation type perception survey, based on the one-question survey, are summarized in Table 5 for human translation and Table 6 for machine translation.

Table 4. Group-Level LMM Results: Differences Between HT and MT

Participant	Variable	Coef	SE	*t*	*p*	LCL	UCL
All (Group)	Intercept	0.334	0.005	68.73	0.000	0.324	0.343
	GFD/ST	0.025	0.007	3.839	0.000	0.012	0.038
	Intercept	0.367	0.004	82.67	0.000	0.358	0.375
	GFD/TT	-0.012	0.006	-2.011	0.044	-0.023	0.000
	Intercept	5.962	1.086	5.487	0.000	3.828	8.095
	GPD/TT	1.054	1.487	0.709	0.478	-1.866	3.976

Table 5. Survey Results for HT

Answer	P1	P2	P3	P4	P5	P6	P7
HT		x		x		x	
Likely HT					x		x
Not Sure	x		x				
Likely MT							
MT							

Table 6. Survey Results for MT

Answer	P1	P2	P3	P4	P5	P6	P7
HT			x			x	
Likely HT		x		x			
Not Sure					x		
Likely MT	x						
MT							x

5. Correlation of Cognitive Effort and Perception

The results reveal varied relationships between cognitive effort and participants' perceptions of whether a translation was produced by a human or a machine. For the source text, participants P4, P5, and P7 exhibited significantly higher fixation durations for MT compared to HT. Interestingly, these participants often misclassified MT as HT in the survey, suggesting that increased cognitive effort does not necessarily lead to accurate identification of translation type. This misclassification may be linked to overcompensation during revision or the perceived fluency of MT. Conversely, P1, who showed significantly lower fixation durations for MT in the target text (FD/TT), accurately identified the HT, suggesting that in some cases, lower cognitive effort during MT tasks may facilitate more accurate classification. However, for most participants, fixation durations in the target text (FD/TT) did not significantly differ between MT and HT, and patterns of perception varied. These findings reinforce the overall inconsistency observed in the individual-level fixation data and suggest that the relationship between cognitive effort and perception is not uniform across participants or text segments.

Pause duration analysis showed minimal differences between MT and HT across participants. These differences were small and not statistically significant, suggesting that pause-based cognitive effort did not consistently align with participants' perceptions. Nevertheless, pause durations provide additional insight into task-specific cognitive demands that may not always be reflected in perception data.

At the group level, MT post-editing required greater fixation effort for the source text, which aligns with the notion of increased processing difficulty. However, MT was often misclassified as HT in the perception survey by several participants, reflecting an inability to link higher cognitive effort with the translation source. For the target text, slightly lower fixation durations for MT were associated with greater uncertainty in responses, suggesting that reduced cognitive effort in MT tasks may contribute to weaker confidence in identifying translation type. Although average pause durations for MT were also longer than those for HT, this difference was not statistically significant, reinforcing the idea that pause-based cognitive effort may be less consistently linked to perception outcomes at the group level.

Survey responses revealed a clear preference for identifying HT, particularly in HT tasks, where most participants selected either "HT" or "Likely HT." In contrast, MT classifications were more varied, with frequent misidentification and higher levels of uncertainty. This pattern suggests that participants tend to associate fluency and quality with HT, even when cognitive effort metrics indicate comparable or greater effort for MT tasks.

6. Conclusion

This study investigated the relationship between cognitive effort and perception during human translation revision (HT) and machine translation post-editing (MT), combining eye-tracking and keylogging data with survey responses to explore how effort metrics correlate with translator judgments and confidence levels regarding the source of the translation (i.e., whether it was translated by a human or through machine translation). It fills a gap in the literature by providing empirical evidence of the correlation between cognitive effort and translators' perceptions during revision and post-editing.

First, the results showed that MT post-editing generally required higher cognitive effort for source texts. In contrast, target texts showed slightly lower effort for MT compared to HT and pause durations did not indicate any consistent patterns across participants.

Second, a correlation between cognitive effort and perception was observed, but it was not consistent across participants or tasks. Participants misclassified MT as HT in cases where fixation durations were higher, which suggests that the MT text may have been perceived as comparable to HT in terms of quality, at least based on the observed correlation. Additionally, in some cases, lower cognitive effort in MT tasks seemed to correspond with increased uncertainty about whether the translation was produced by a human or a machine, suggesting reduced confidence.

Third, the results are generally consistent with findings reported in prior research (e.g., Daems & Macken, 2020; Robert et al., 2022), more specifically regarding the role of assumptions in shaping performance. Participants' expectations about whether the target text was HT or MT influenced their behavior and ultimate performance (Briva-Iglesias & O'Brien, 2024). This study extends previous work by exploring the correlation between cognitive effort, misclassification, and confidence during the revision and post-editing of translated texts. It adapted eye-tracking and keylogging to provide quantifiable evidence of cognitive effort and used a survey to measure misclassification and confidence. These findings may point to the need for greater awareness of how fluency and task assumptions can shape translator judgment, particularly during post-editing.

Funding: This research received no external funding.
Conflicts of Interest: The authors declare no conflict of interest.

References

[1] Alves, F., Pagano, A., & da Silva, I. (2009). A new window on translators' cognitive activity. In I. Mees, F. Alves, & S. Göpferich (Eds.), *Methodology, technology and innovation in translation process research* (pp. 267–291). Samfundslitteratur.

[2] Associated Press. (2023, November 14). Horse escapes stall on flight from New York to Belgium. AP News. Retrieved January 12, 2024, from https://apnews.com/article/horse-loose-cargo-plane-returns-jfk-airport-6066bedc8438d69fd783fc4516bcf4e0

[3] Briva-Iglesias, V., & O'Brien, S. (2024). Pre-task perceptions of MT influence quality and productivity: The importance of better translator-computer interactions and implications for training. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1).* European Association for Machine Translation.

[4] Carl, M. (2012). Translog-II: A program for recording user activity data for empirical reading and writing research. In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12) (pp. 4108–4112). European Language Resources Association (ELRA).

[5] Daems, J., & Macken, L. (2020). Post-editing human translations and revising machine translations: Impact on efficiency and quality. In M. Koponen, B. Mossop, I. Robert, & G. Scocchera (Eds.), *Translation revision and post-editing* (pp. 50–70).

[6] Daems, J., Vandepitte, S., Hartsuiker, R. J., & Macken, L. (2017). Identifying the machine translation error types with the greatest impact on post-editing effort. *Frontiers in Psychology*, 8, 1282. https://doi.org/10.3389/fpsyg.2017.01282

[7] Doherty, S. (2016). Translations | The impact of translation technologies on the process and product of translation. International Journal of Communication, 10, 947–969.

[8] GP3. (2025). *GP3 Eye Tracker Placement and Positioning*. https://gazept.com/dl/GP3_Placement_and_Positioning.pdf

[9] Jakobsen, A. L. (2018). Moving translation, revision, and post-editing boundaries. In H. V. Dam, M. N. Jensen, & J. Ø. Hansen (Eds.), *Moving boundaries in translation studies* (pp. 64–80). Routledge.

[10] Jakobsen, A. L. (2017). Translation process research. In J. W. Schwieter & A. Ferreira (Eds.), *The handbook of translation and cognition* (pp. 19–49). Wiley-Blackwell.

[11] Koponen, M. (2016). Is machine translation post-editing worth the effort? A survey of research into post-editing and effort. The Journal of Specialised Translation, 25, 131–148.

[12] Nitzke, J., & Gros, A.-K. (2020). Preferential changes in revision and post-editing. In M. Koponen, B. Mossop, I. Robert, & G. Scocchera (Eds.), *Translation revision and post-editing* (pp. 21–34). Routledge.

[13] Pym, A. (2013). Translation skill-sets in a machine-translation age. *Meta, 58*(3), 487–503. https://doi.org/10.7202/1025047ar

[14] Robert, I. S., Ureel, J. J. J., & Schrijver, I. (2022). Translation, translation revision and post-editing competence models: Where are we now? In G. Massey, E. Huertas-Barros, & D. Katan (Eds.), *The human translator in the 2020s* (pp. 44–59). Routledge. https://doi.org/10.4324/9781003223344

[15] Rydning, A. F., & Lachaud, C. M. (2010). The Reformulation Challenge in Translation. In *Translation and Cognition* (pp. 85–108).

[16] Saldanha, G., & O'Brien, S. (2014). Research Methodologies in Translation Studies. Routledge.