Çankaya University Journal of Humanities and Social Sciences
2025, Vol. 19, no. 2, 293-309
https://doi.org/10.47777/cankujhss

Translation Studies | research article

Effort in Machine Translation Post-Editing: The Role of Expertise

Makine Çevirisi Sonrası Düzenlemede Çaba: Uzmanlığın Rolü

Minel Sayar Öztürk 0000-0001-7015-4978
Alper Kumcu 0000-0003-0844-3562
Hacettepe Üniversitesi

ABSTRACT
Machine translation post-editing is becoming an increasingly important part of the translation process, boosting efficiency. As the primary aim of technology is to reduce human effort, the concept of effort has been widely investigated in this context. This study examines the relationship between expertise and effort in MTPE, addressing the following two questions: (1) To what extent can expertise serve as a valid indicator of cognitive, temporal, and technical/linguistic effort in MTPE? (2) What is the relationship between expertise, objective and self-reported effort, and post-editing accuracy? A mixed-methods experiment was conducted with 21 participants divided into three groups: experienced translators, inexperienced translators, and field experts. They were asked to post-edit three legal documents that had been machine-translated from English to Turkish. Their post-editing processes were recorded using keylogging software (Translog-II), which measured total task duration, mean pause duration, total pause duration, pause percentage, pauses per word, text production and elimination, and user and production events per minute. Post-editing accuracy, self-reports and retrospective think-aloud records were also collected. The findings revealed that experienced translators exerted higher cognitive and temporal effort than inexperienced translators, as well as higher technical and linguistic effort than inexperienced translators and field experts combined.

ARTICLE HISTORY
Received 12 Sept 2025
Accepted 23 Oct 2025

KEYWORDS
Machine translation post-editing (MTPE), cognitive effort, temporal effort, technical effort, expertise

Introduction

Recent technological advancements and the immense increase in business volume driven by globalisation and the circulation of knowledge have brought about a shift in the roles of practitioners and the medium they use to fulfil their tasks. Technological tools and computerised systems have therefore become both inevitable and indispensable in translation practice, although human involvement remains preferable. Machine translation (MT) is one of the most frequently used technologies in translation. It is widely adopted by users for purposes such as communication, assimilation, and dissemination (see Hutchins 2001, for a discussion) and is among the primary tools utilised by translation practitioners. Many of these practitioners argue that MT accelerates the completion of translation tasks (Paradowska et al. 2025). Although neural machine translation (NMT), an advanced form of MT that mimics a human neural network through

© 2025 The Author(s). CUJHSS (e-ISSN 3062-0112) Published by Çankaya University.
Minel Sayar Öztürk, PhD candidate, Dept. of Translation and Interpreting Studies, Hacettepe University, Turkiye | minelsayarms@gmail.com; ORCID# 0000-0001-7015-4978; https://doi.org/10.47777/cankujhss
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

deep learning (Sharma et al. 2023, pp. 2-4), offers significant advantages in producing well-structured and fluent target texts, it still requires the application of post-editing (PE), particularly for dissemination purposes. This is because MT output often contains translation errors ranging from grammatical to stylistic, which must be corrected to ensure quality (O'Brien, 2022, p. 106). Machine translation post-editing (MTPE) is defined as a dynamic process that places the PE practitioner at the centre of the translation workflow and continues until the most accurate version of the target text is achieved (Rico Pérez 2024, p. 32).

Various studies have examined the relationship between PE and cognitive effort at the intersection of translation studies and cognitive science (e.g., Herbig et al. 2021; Lacruz 2017; Lacruz et al. 2012; O'Brien 2006). Krings (2001) categorises PE effort into three types: temporal effort, cognitive effort, and technical effort. While it is possible to obtain precise data on temporal effort by directly measuring the time spent on a task (O'Brien 2011), and on technical effort by recording physical actions; through keylogging, for instance (Koponen et al. 2020), there is still no single and direct method for collecting data on cognitive effort (Lacruz et al. 2012). As a result, assessing cognitive effort requires the combination of multiple data sources to draw meaningful conclusions about the mental workload exerted by the PE practitioner. In several studies, pauses have been accepted as a reliable metric for cognitive effort, which is considered the most challenging type of effort to measure (Lacruz 2017, p. 387). For example, Koponen et al. (2019) conducted a case study comparing the PE alterations of neural, rule-based, and statistical machine translation (MT) systems for the Finnish-English language pair, employing both product- and process-oriented approaches. They found that NMT output generally required less technical effort to edit, while RBMT output demanded more cognitive effort. However, the mixed-system experimental design limited the ability to link effort indicators to specific outputs. Lacruz et al. (2012) introduced the average pause ratio as a potential measure of cognitive demand, revealing that more cognitively demanding segments had a smaller pause ratio. However, they acknowledged that their study involved only one participant and emphasised the need for replication with larger samples to ensure generalisability.

Studies that adopt specific pause durations to identify cognitively demanding pauses show considerable variation in researchers' preferences due to methodological differences and diverse study objectives (Kumpulainen 2015, p. 49). For example, in his PhD research on the German-French language pair, Krings (1986) defined cognitively demanding pauses as those lasting more than three seconds, whereas Lörscher (1991) opted for pauses lasting more than two seconds for the same language pair. Dragsted et al. (2009) focused on pauses of more than 2.5 seconds in their research on the Danish-English language pair, whereas Jakobsen (2003) used a threshold of five seconds. Similarly, Yanmei and Mingmei (2014) examined pauses of over five seconds in their study of Chinese-English language pairs. In contrast, Immonen (2006) examined both extremely short pauses of 0.01 seconds and longer pauses in the Finnish-English language pair. However, Dragsted (2004) and Mutta (2016), highlighting the unique pause behaviour of each participant in experimental studies, rejected the use of a single pause threshold, instead adopting individual-specific pause durations.

In terms of the perceptions of cognitive effort and cognitive load held by translators and interpreters, an empirical study (Gieshoff & Heeb 2023) shows no clear distinction between the two. However, the literature typically defines them as follows: Cognitive load refers to the inherent difficulty of the task, while cognitive effort reflects the mental exertion required to complete it (Ehrensberger-Dow et al. 2020). The present research focuses on the cognitive effort exerted by participants when completing MTPE tasks involving comparable levels of cognitive load. It aims to identify the group best suited to MTPE by categorising participants based on their expertise. According to Muñoz Martín (2014), translation expertise encompasses five situated dimensions:

knowledge, referring to declarative or conscious information stored in long-term memory; adaptive psychophysiological traits, which are skills essential for enhancing performance; problem-solving skills, involving the ability to analyse issues, identify solutions, and apply effective strategies; regulatory skills, meaning metacognitive abilities that enable translators to monitor, control, and adjust their cognitive processes; and self-concept, which relates to self-awareness, perception of the environment, and confidence (pp. 18-28). A deficiency or inadequacy in these abilities creates a gap in translation expertise that requires a multidimensional skill set. However, these skills alone are insufficient, particularly for translators who specialise in specific fields, such as legal translation. As Šarčević (1997) points out, legal translators require a combination of key skills, including knowledge of legal terminology and an understanding of the legal systems of both the source and target languages. They must also be able to address legal issues and demonstrate proficiency in analysis, forecasting, and drafting within a legal context (pp. 113-114). This suggests that expertise comprises both the fundamental skills necessary for producing accurate translations and the specialised competencies required for professional performance in domain-specific areas of translation.

Present Study

The brief review of the studies above suggests that keylogging is one of the most reliable methodologies for measuring effort in the translation process, and in particular, pauses indicate cognitive effort. Furthermore, expertise is a phenomenon that might influence effort through the set of abilities it naturally brings to translation tasks. Against this background, the present study focuses on MTPE, one of the primary applications in contemporary translation workflows, which is a result of the profession's evolving nature. In general, the present study aims to investigate the effect of different types of expertise on effort, with a specific focus on cognitive effort, in MTPE through an experiment involving experienced translators, inexperienced translators, and field experts. We aim to provide insights into MTPE in the English-Turkish language pair and contribute to the field of legal translation by examining MTPE within the context of the legal genre. In particular, the study seeks to address the following research questions: (1) To what extent can expertise serve as a valid indicator of cognitive, temporal, technical, and linguistic effort in MTPE? (2) What is the relationship between expertise, objective and self-reported effort, and post-editing accuracy?

Method

Participants

Participants (n = 21, M_{age} = 32.9, SD = 7.6) were selected using the convenience sampling method with purposive subgrouping and divided into three subgroups as experienced translators (n = 7, M_{age} = 37.4, SD = 6.3), inexperienced translators (n = 7, M_{age} = 28.9, SD = 2.5), and field experts (n = 7, M_{age} = 32.4, SD = 10.1). The experienced translator group consisted of individuals with proven expertise in legal translation. In contrast, the inexperienced group included participants who were actively engaged in translation activities, either professionally or academically, but had little to no prior experience with legal texts. The field experts, on the other hand, were professionals from the legal domain, such as practising lawyers or academics in law and law-related disciplines, but with little to no translation experience. All participants were native Turkish speakers who used it as their primary language of communication. English was their second language.

According to the pre-experiment questionnaire, the experienced translator group consisted of three participants with a bachelor's degree, three with a master's degree, and one with a doctoral degree. In the inexperienced translator group, two participants held bachelor's degrees, and five held master's degrees. The field expert group consisted of three participants with a bachelor's degree, three with a master's degree, and one with a doctoral degree. In other words, the

participants' educational backgrounds were similar. In terms of professional identity, six of the experienced translators identified solely as translators, while one identified as both a translator and an academic. Among the inexperienced translators, one participant identified solely as a translator, five identified as academics specialising in translation studies, and one identified as other, being employed as a programme associate at an international organisation. Of the field experts, four identified as lawyers, one as an academic, one as both a lawyer and an academic, and one as both a translator and an academic. The academics in the field experts group worked in law or law-related disciplines. All participants, on the other hand, were fully informed about the procedures and provided written informed consent. The study was approved by the Social Sciences Ethics Committee of Hacettepe University (approval number: E-66777842-300-00003055757).

Material

To determine the genre of textual material to be used in the study, pre-experiment field research was conducted with five large-scale translation bureaus in Ankara, Türkiye. The research suggested that legal texts represent a significant portion of the translation industry in Türkiye. Consequently, this genre was selected to represent the Turkish market in the experiment. Accordingly, three legal agreements in English were developed for the experiment and then translated into Turkish to create the target texts. The texts were translated using DeepL, the most frequently used MT tool in the pilot study. Descriptive analyses were performed for both the source and target texts to ensure comparability. The readability of the English texts was measured using the Flesch–Kincaid Readability Index, while the Turkish translations were assessed using the Ateşman reading ease formula. Additional metrics, such as average sentence length and type-token ratios, were also calculated. As shown in Table 1, all three English source texts were classified as "very difficult" according to the Flesch–Kincaid Readability Index (Kincaid et al. 1975), with an average of 24.0 - 26.15 words per sentence. The corresponding Turkish translations also fell within the "difficult" or "very difficult" categories according to the Ateşman Reading Ease Index (Ateşman 1997), featuring slightly shorter sentences and higher syllabic complexity (see Table 1). This is to be expected, given the linguistic and structural characteristics of legal texts as a genre. Type-token ratios ranged from 0.44 to 0.49 in the source texts and from 0.59 to 0.63 in the target texts. These ratios indicate a high level of lexical variety across all materials. Overall, these findings suggest that the source and target versions of the three texts are comparable in terms of difficulty and lexical richness, ensuring consistency in the materials used for the experiment.

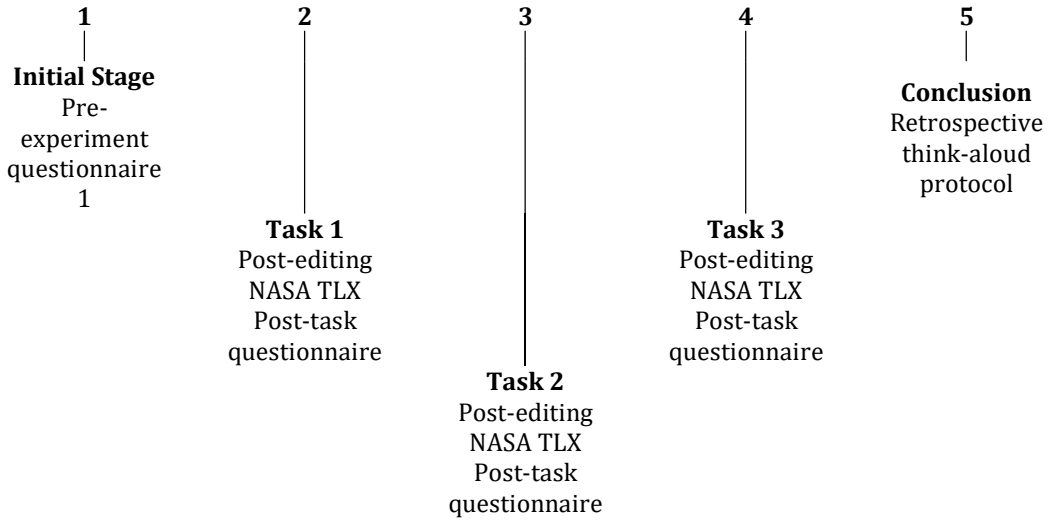
Table 1
Descriptive Statistics for the Source and Target Texts Used in the Experiment

Measure	Alpha (EN)	Alpha (TR)	Macro (EN)	Macro (TR)	Global (EN)	Global (TR)
Word count	494	415	523	462	552	489
Sentence count	20	20	20	20	23	22
Avg. words per sentence	24.7	20.75	26.15	23.1	24.0	22.23
Avg. syllables per word	1.96	2.84	1.95	2.86	2.04	2.98
Readability score	11.39	30.63	-6.18	23.75	-9.35	21.27
Readability category	Very Diff.	Difficult	Very Diff.	Very Diff.	Very Diff.	Very Diff.
Types	215	248	249	283	263	297
Tokens	491	424	510	449	536	472
Type–token ratio	0.44	0.59	0.49	0.63	0.49	0.63

Procedure

A pilot study was conducted before the main experiment to evaluate and refine the research design. The pilot study involved one individual from each of the three participant groups, none of whom participated in the actual experiment. The actual experiment comprised five steps (see Figure 1):

Figure 1
Experimental Procedure



(1) All participants were asked to complete a pre-experiment questionnaire on Google Forms. The questionnaire consisted of seven questions on demographics, professional background and translation experience.

(2) Participants were asked to post-edit the MT output in Translog-II (Carl 2012), with their keyboard and mouse movements being recorded by the software. In Translog-II, the original English text appeared on the left of the screen, with the Turkish machine translation on the right (see Figure 2). Thus, participants worked on the MT output displayed in the right-hand window. As the source texts were in English and the MT outputs were in Turkish, the PE tasks were also in Turkish, which was the participants' first language. Beforehand, they were provided with written instructions on how to use Translog-II, and their queries were addressed. They were explicitly informed that the "select all" (Ctrl+A) and "undo" (Ctrl+Z) keyboard shortcuts were not functional in the software. Since three textual materials were used in the experiment as specified above, the MTPE task was repeated three times for all participants. In addition to the visual view of the MTPE process that revealed the pause-related metrics, Translog-II recorded statistics on total user events, text production, text elimination, total duration, user events per minute and text production per minute. These metrics were used in the following analyses.

(3) After each of the three MTPE tasks, all participants were asked to complete a post-task questionnaire via Google Forms. The questionnaire consisted of four 5-point Likert-scale questions and one open-ended question. This was designed to evaluate participants' perceptions of the quality of the MT output and their self-assessed performance.

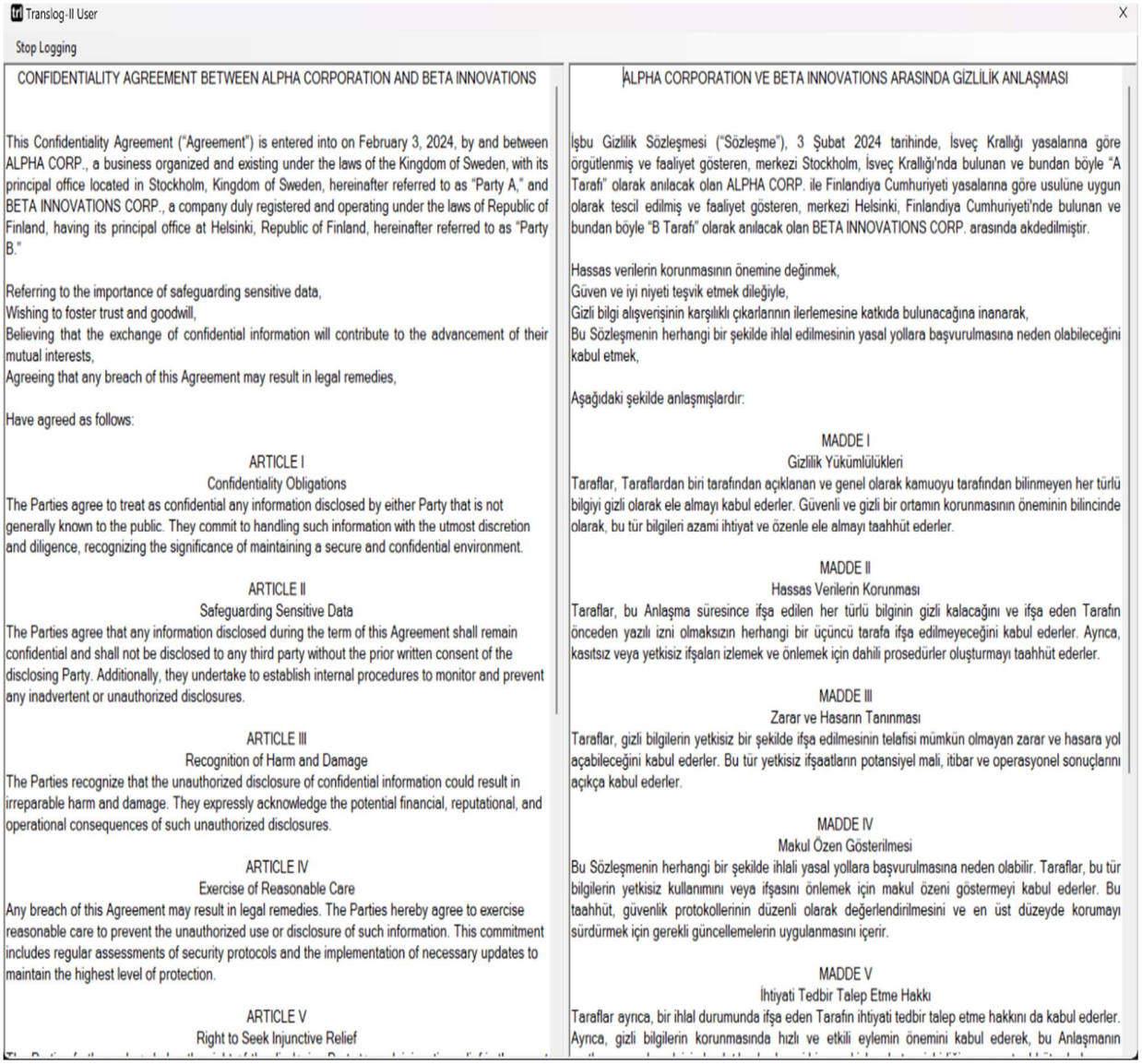
(4) Participants were asked to complete a paper-based NASA Task Load Index (NASA-TLX) (Hart & Staveland 1988). The NASA-TLX consists of six workload dimensions, each rated on a scale of 0 to 21. Accordingly, participants were asked to evaluate the mental, physical and temporal demands

of the task, as well as their performance, effort and frustration levels. Steps 2–4 were repeated for each of the three MTPE tasks.

(5) Once all three MTPE tasks and their associated assessments had been completed, a retrospective think-aloud protocol (rTAP) was carried out. Participants answered open-ended questions about their PE experience and the effort they had put into the process, while their responses were auditorily recorded. The aim was to shed light on the rationale behind certain MTPE decisions that cannot be explained solely by numbers, and to elucidate the cognitive processes underlying the quantitative data collected using the Translog-II.

Questionnaires and rTAP questions in Turkish can be accessed via the project's data repository at https://osf.io/a5ktn.

Figure 2
Interface of Translog-II during an MTPE task



For logistical reasons, the experiment was conducted in two laboratories, enabling participants to choose the most accessible setting. The labs were the Postgraduate Research Laboratory at the Department of Translation and Interpreting at Hacettepe University and the Department of Translation and Interpreting at Çankaya University. At Hacettepe University, participants used a desktop computer with a 1600 x 900 resolution monitor, an MSI-brand keyboard, and a mouse. At Çankaya University, the monitor resolution was 1366 x 768. In other words, the hardware specifications in both settings were comparable. Similarly, the room arrangements were standardised to feature a comfortable chair, a large desk, and minimal distractions. Both labs also provided consistent environmental conditions, featuring bright but non-glaring overhead lighting and quiet settings free from external noise, thereby ensuring that light and sound did not interfere with participants' performance.

Data Analysis

The PE tasks were recorded using Translog-II (Carl 2012). Translog-II is a computer-based keylogging tool widely used in research on the translation process. It records translators' keystrokes, mouse movements, and pauses in real-time while they work on a translation or PE task. The software also provides playback and visualisation features, enabling researchers to analyse writing and revision behaviour, processing time and cognitive effort during translation. Translog-II is particularly useful for examining the dynamics of the translation process, such as pauses, text production, and editing patterns. Researchers can customise the page displayed during the experiment by providing the source and target texts and configuring alignment and font settings.

The following measures were collected for each participant across three tasks: total task duration (seconds), mean pause duration (seconds), total pause duration (seconds), pause percentage (%), pauses per word, text production (number of characters), text elimination (number of characters), user events (count), user events per minute and production events per minute. These keylogging metrics can be systematically mapped onto the effort types in MTPE (temporal, cognitive, technical and linguistic). Accordingly, the total task duration was recorded in seconds to measure the overall time spent on each task, indicating temporal effort. Pause-related measures, such as mean pause duration, total pause duration, pause percentage and pauses per word, were also measured to indicate cognitive effort during MTPE. Pauses lasting more than two seconds were selected for analysis, as this threshold corresponds with the average duration observed in previous studies (e.g., Lörscher 1991). These pauses were identified using the software's configuration options and exported to an Excel spreadsheet for further analysis. Text production, text elimination, user events, user events per minute and production events per minute, on the other hand, provided insights into technical and linguistic effort. In this context, an event refers to any user action recorded in Translog-II, including keystrokes, deletions, and cursor movements.

The data were analysed and visualised using the R programming language, version 4.5.1 (R Core Team 2025). Descriptive data were analysed and presented with the psych package (Revelle 2018). A total of 63 MPTE sessions were recorded. The data were then used to conduct group comparisons among experienced translators, inexperienced translators, and field experts using one-way ANOVA tests. Findings from NASA-TLX concerning mental demand, physical demand, temporal demand, performance, effort, and frustration for each task were analysed to predict effort metrics by fitting linear regression models.

Lastly, to obtain more comprehensive results, the post-edited texts were evaluated for accuracy. Each sentence was scored as follows: Two points were awarded for accurate translations, one point for partially accurate ones, and zero points for mistranslations or omissions. After scoring, normative scores were calculated for each of the 63 texts and recorded in an Excel spreadsheet for

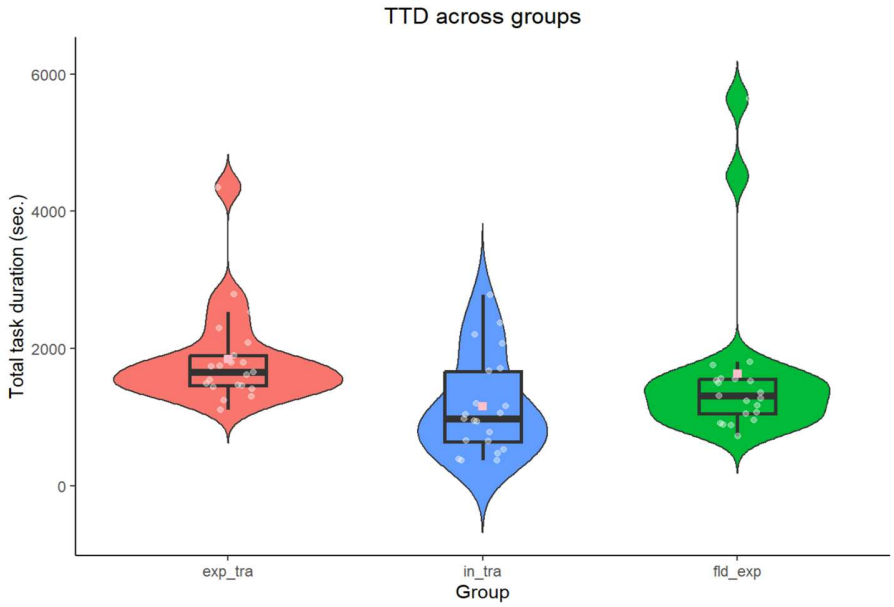
further analysis in R. This enabled comparisons to be made across participant groups. Raw and processed data can be accessed via the project's data repository at https://osf.io/a5ktn.

Results

Temporal Effort

Total task duration (in seconds), as an indicator of temporal effort in MTPE, was analysed across three different participant groups (i.e., experienced translators, inexperienced translators, and field experts). Descriptive analysis revealed that experienced translators spent the longest average time, in seconds, across three MTPE tasks (M = 1847.11, SD = 710.52), followed by field experts (M = 1629.88, SD = 1198.23) and inexperienced translators (M = 1158.41, SD = 711.78). The distributions for the experienced translators (skewness = 2.10, kurtosis = 4.67) and field experts (skewness = 2.37, kurtosis = 4.63) were notably right-skewed with heavy tails, reflecting the presence of extremely high values. In contrast, the inexperienced translators showed a more balanced distribution (skewness = 0.78, kurtosis = - 0.62). A one-way ANOVA revealed a statistically significant difference in total task duration among the three groups, $F(2, 60) = 3.19$, p = .048. Post-hoc Tukey tests indicated that the experienced translators completed the task significantly slower than the inexperienced translators; mean difference = 688.70, 95% *CI* [1358.55, 18.85], p = .043. No significant differences were found between the field experts and either of the other two groups (all ps > .05). These patterns are also illustrated in Figure 3, where violin plots depict the distribution, spread, and median total task durations for each group.

Figure 3
Total Task Duration across Groups

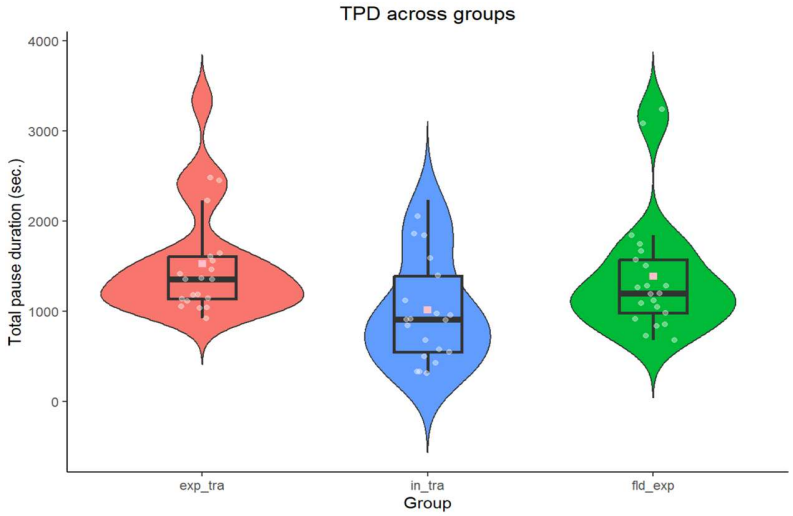


Cognitive Effort

Total pause duration (in seconds) and pauses per word (count), as indicators of cognitive effort in MTPE, were analysed across three different participant groups (i.e., experienced translators, inexperienced translators, and field experts). Descriptive statistics indicated that experienced translators had the longest average total pause duration in seconds (M = 1528.92, SD = 609.00), followed by field experts (M = 1388.33, SD = 673.68) and inexperienced translators (M = 1015.65, SD = 595.28). The distributions for the experienced translators (skewness = 1.49, kurtosis = 1.46) and field experts (skewness = 1.59, kurtosis = 1.82) were positively skewed with some high values,

indicating the presence of participants who paused for substantially longer durations. In contrast, the inexperienced translators displayed a more balanced distribution (skewness = 0.63, kurtosis = - 0.92). A one-way ANOVA revealed a statistically significant difference in total pause duration among the three groups, $F(2, 60) = 3.76$, $p = .029$ (see Figure 4). Post-hoc Tukey tests indicated that the experienced translators paused significantly longer than the inexperienced translators; mean difference = 513.28, 95% CI [978.23, 48.32], $p = .027$. No significant differences were found between the field experts and either of the other two groups (all *p*s > .05).

Figure 4
Total Pause Duration across Groups

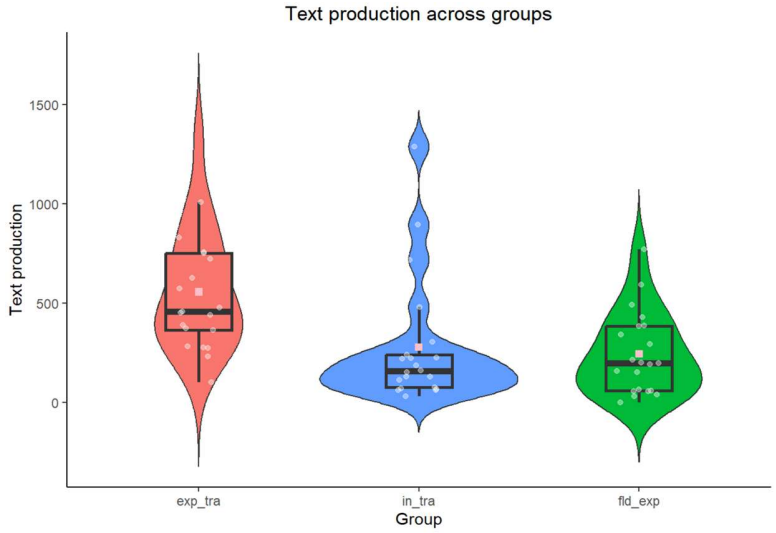


We also analysed pauses per word across groups. Descriptive statistics indicated that the experienced translators recorded the highest pause per word ($M = 3.37$, $SD = 1.34$), followed by the field experts ($M = 3.03$, $SD = 1.39$) and the inexperienced translators ($M = 2.23$, $SD = 1.30$). The experienced translators (skewness = 1.45, kurtosis = 1.31) and field experts (skewness = 1.57, kurtosis = 1.62) both exhibited positively skewed distributions with heavier tails, indicating that some participants had higher pause per word counts. In contrast, the distribution of inexperienced translators was more balanced (skewness = 0.63, kurtosis = - 0.88). A one-way ANOVA revealed a statistically significant difference in pauses per word among the three groups, $F(2, 60) = 3.94$, $p = .025$. Post-hoc Tukey tests indicated that the experienced translators had significantly more pauses per word than the inexperienced translators; mean difference = 1.14, 95% CI [2.13, 0.14], $p = .022$. No significant differences were found between the field experts and either of the other two groups (all *p*s > .05).

Technical and Linguistic Effort

Text production (in terms of character count), text elimination (in terms of character count), and production events per minute, as indicators of technical and linguistic effort in MTPE, were analysed across three different participant groups (i.e., experienced translators, inexperienced translators, and field experts). For text production, a one-way ANOVA revealed a statistically significant difference among the three groups, $F(2, 60) = 7.83$, $p < .001$ (see Figure 5). Post-hoc Tukey tests indicated that the field experts produced significantly shorter text than the experienced translators; mean difference = 313.10, 95% CI [521.69, 104.51], $p = .002$, and the inexperienced translators also produced significantly shorter text than the experienced translators (mean difference = 278.90, 95% CI [487.49, 70.31], $p = .006$). No significant difference was observed between the field experts and the inexperienced translators ($p = .918$).

Figure 5
Text Production across Groups



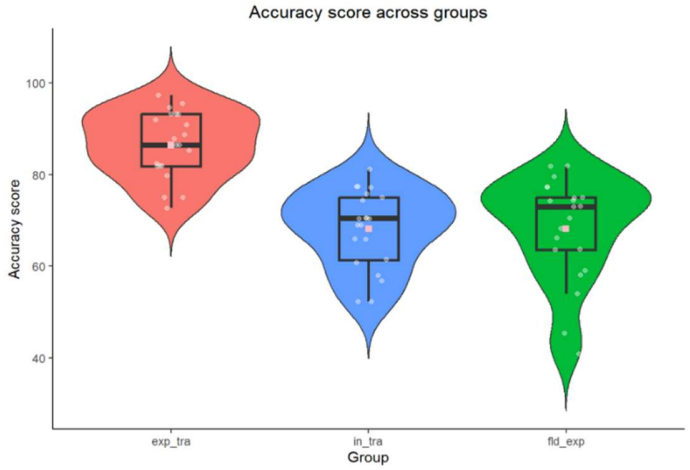
Similarly, a one-way ANOVA revealed a statistically significant difference in text elimination among the three groups, $F(2, 60) = 3.72$, $p = .030$. Post-hoc Tukey tests revealed that the inexperienced translators eliminated significantly fewer characters than the experienced translators; mean difference = 135.29, 95% CI [13.03, 257.54], $p = .027$. No significant differences were found between the field experts and either of the other two groups (all *p*s > .05).

A one-way ANOVA revealed a statistically significant difference in production events per minute among the three groups, $F(2, 60) = 5.26$, $p = .008$. Post-hoc Tukey tests indicated that the field experts produced significantly fewer events per minute than the experienced translators; mean difference = 9.73, 95% CI [16.95, 2.52], $p = .005$. No significant differences were found between the inexperienced translators and either of the other two groups (all *p*s > .05).

Accuracy

A one-way ANOVA revealed a statistically significant difference in accuracy scores among the three groups, $F(2, 60) = 27.73$, $p < .001$ (see Figure 6). Post-hoc Tukey tests revealed that experienced translators achieved significantly higher accuracy scores than both field experts (mean difference = 18.21, 95% *CI* [11.41, 25.00], $p < .001$) and inexperienced translators (mean difference = 18.25, 95% *CI* [11.46, 25.05], $p < .001$). However, no significant difference was found between the field experts and the inexperienced translators ($p > .05$).

Figure 6
Accuracy Score across Groups



Prediction Models

Linear regression models were used to investigate how well key effort measures reported above predicted demand, as measured by NASA-TLX. The models revealed that total task duration; B = 139.71, $t(61) = 3.32$, $p = .002$, pauses per word; $B = 0.23$, $t(61) = 3.76$, $p < .001$, and total pause duration; $B = 110.18$, $t(61) = 3.83$, $p < .001$ was a significant predictor of physical demand. Consequently, higher physical demand scores were associated with longer task durations, more pauses per word and longer total pause durations. In other words, as total task duration, pauses per word and total pause duration increased, participants experienced and reported higher levels of physical demand, but not of other types of demand.

Retrospective Think-aloud Protocol

rTAP provided a platform for participants to share their reflections on the post-editing process. The questions in the protocol were adapted from a study by Ruokonen and Koskinen (2017), which used reflective writing tasks to explore translators' emotional narratives and perceptions of technology. While the original study focused on how translators described the agency of machines, the questions in this study were rephrased to encourage participants to share their experiences of the post-editing task they had completed. During the rTAP, participants were asked about their general impressions of the MT outputs they encountered in the experiment, the reasons behind the pauses observed during the MTPE process, and the types of support they found most crucial and would be most reluctant to lose. Participants were also invited to reflect on the most frustrating hindrance they had encountered and the aspect they would most like to eliminate, and to share any additional comments they wished to make.

Some participants explicitly expressed positive views of the material to be edited, emphasising that the MT outputs were adequate and of good quality, saving them time. In contrast, others raised concerns about terminological inconsistencies, the improper use of Turkish, and the clear need for MTPE. Interestingly, these positive and negative opinions were expressed by participants across all groups, with no particular group exhibiting a greater prevalence of either. In other words, participants' level of expertise did not influence their evaluation of the MT output.

That said, most experienced translators emphasised the need for PE of MT outputs, whereas those who said that MT facilitated the translation process were mainly inexperienced. One experienced translator explicitly stated that MT outputs increased the time and cognitive effort required, while two others indicated that PE was more complex than translating from scratch. While some

experienced translators and field experts considered the terminology in MT outputs to be insufficient, several inexperienced translators and field experts were satisfied with the terminological choices provided by the MT tool. Two field experts noted that the three source texts represented different legal subgenres and that their expertise varied across them since it is not possible to specialise in every legal subgenre.

Most experienced translators and field experts indicated that the main reasons for pausing during MTPE tasks were to read the source and target texts and identify potential errors. They particularly emphasised the need for time to process long, complex legal sentences. One experienced translator explained that they paused to detect recurring mistakes and, therefore, occasionally stopped to scan the entire passage. Conversely, a field expert noted that they paused to check conjunctions and transitions, emphasising their critical importance in the legal context. By contrast, some inexperienced translators reported physical discomfort, such as stomachaches and eyestrain, as reasons for pausing during MTPE tasks.

Lastly, one inexperienced translator stated that they ignored the source text displayed on the screen, focusing only on the target text to be post-edited. Another inexperienced translator explicitly indicated that they preferred to preserve the MT output as much as possible, believing that the MT tool produced higher-quality results than they could.

Discussion

The present study examined effort in MTPE and the role of expertise in this context. Specifically, we investigated the extent to which expertise in MTPE tasks can be considered a valid indicator of cognitive, temporal, technical, or linguistic effort, as well as the relationship between expertise, objective and self-reported effort, and PE accuracy. To this end, a mixed-methods design combining qualitative and quantitative data through triangulation was employed. Data were collected from 21 participants, each of whom completed three MTPE tasks, resulting in a total of 63 data points. The dataset comprised behavioural measures consisting of total task duration (in seconds), total pause duration (in seconds), mean pause duration (in seconds), pause percentage, pause per word ratio, text production (number of characters), text elimination (number of characters), user events count, user events per minute count, and production events per minute count, as well as self-reports based on the NASA-TLX dimensions of mental demand, physical demand, temporal demand, performance, effort, and frustration. Qualitative data were also obtained through rTAP interviews to provide further insights into the MTPE process. Accuracy scores were calculated from the MT post-edited texts produced by the participants to provide an additional measure of effort in MTPE.

Analysis of total task duration, an indicator of temporal effort, revealed that experienced translators spent more time on MTPE than their inexperienced counterparts. In other words, they exerted more temporal effort than the inexperienced translators. Analysis of total pause duration and pause per word, both of which are indicators of cognitive effort, revealed that experienced translators paused for longer overall and for a longer duration per word than inexperienced translators. This suggests that experienced translators exerted higher cognitive effort than inexperienced translators. Finally, analyses of text production and elimination, as well as production events per minute, taken together as indicators of technical and linguistic effort, demonstrated that experienced translators produced and omitted more text, and generated more production events per minute, than inexperienced translators and field experts. In other words, experienced translators demonstrated the greatest technical and linguistic effort compared to both groups. The keylogging metrics made it critical to refer to accuracy scores to contextualise and interpret the findings. The accuracy scores of the 63 post-edited texts indicated that translation expertise was the decisive factor in achieving high-quality MTPE output. Experienced

translators significantly outperformed inexperienced translators and field experts, who did not differ from each other.

The results of the behavioural measures clearly showed that experienced translators invested more temporal, cognitive, and technical/linguistic effort in MTPE. However, their superior performance was also evident in their accuracy scores, as they produced the most accurate post-edited versions of the MT outputs, outperforming both inexperienced translators and field experts. In other words, although inexperienced translators and field experts exerted less temporal, cognitive, and technical/linguistic effort, their general linguistic ability, without domain-specific training, did not yield qualified MTPE outcomes. Similarly, the domain knowledge possessed by field experts was insufficient to ensure accuracy. rTAPs shed light on the possible grounds behind the higher temporal, cognitive, and technical/linguistic effort invested by experienced translators during the MTPE tasks. An experienced translator indicated that the reason for the initial long pause was to read the entire passage first to analyse it before proceeding segment by segment to make the necessary changes. Some experienced translators and field experts reported that the texts were complex due to long sentence structures and dense terminology, which are typical of legal texts, and that they occasionally paused to read and understand the content. Furthermore, experienced translators, in particular, found the MT output to be inadequate in terms of terminology. Correspondingly, Popovic et al. (2014) demonstrated that the highest cognitive effort was required for lexical and word order editing, influencing PE time. In the same vein, Sánchez-Torrón et al. (2025) stated that deep learning systems may produce terminology that appears accurate, yet is not always reliable, particularly in domain-specific fields such as legal texts, as in this study. They also emphasised the importance of human expertise in validating terminological choices. Similarly, Ibrahim (2025), in a comparison of English–Arabic translation accuracy, suggests a hybrid approach combining MT and human agency to save time and improve translation quality, again highlighting the indispensable role of human expertise. Based on previous studies as well as the objective and subjective data obtained, our findings also support the significance of "human touch" in MT outputs to ensure equivalent translation. It should also be acknowledged that when experienced translators edited the MT output, this increased the technical and linguistic effort involved, as the process included text elimination, text production and user events. This may explain why technical/linguistic effort was the highest among experienced translators as well. Similarly, these properly post-edited terms during the MTPE process may have contributed to the highest accuracy scores among the groups.

When the objective keylogging measures and PE accuracy are considered alongside the subjective data obtained from the rTAPs, it becomes clear that the experienced translators worked more meticulously to finalise the MT outputs than the other groups. This behaviour can be interpreted as an outcome of their translation expertise. In this regard, Koponen (2016) found that MTPE can be beneficial under certain conditions, particularly when the translator's experience and individual work processes are considered. Consistent with this, the present findings suggest that experienced translators were better able to detect and correct errors when they adopted a careful and deliberate approach to the texts, benefiting from the MTPE system in the process. However, it should be underlined that although expertise is often thought to reduce the time needed to complete tasks in fields such as translation (Dong & Chen 2025; Weng & Zheng 2024), medicine and surgery (Persky & Robinson 2017) and forensic fingerprint analysis (Robson et al. 2021), our findings suggest otherwise. This may be because the specific set of competencies required for PE distinguishes it from translation practice more broadly. In addition to linguistic skills, effective PE requires digital literacy to work with MT tools, the ability to assess risk when evaluating MT outputs, and the capacity to anticipate errors to detect and correct problematic segments (Eszenyi 2025). Rico Pérez (2024) has proposed that MTPE should be understood as part of a broader ecosystem that encompasses linguistic correction as well as cognitive, technological and

professional dimensions. In this ecosystem, the process is shaped by the expertise of translators, digital tools, working conditions, and institutional expectations. These factors encourage a meticulous approach, involving careful analysis, error detection, and correction, to produce the best possible MT output. This is demonstrated by the experienced translators in this study.

According to linear regression models based on NASA-TLX findings, physical demand was found to be a significant predictor of total task duration, pauses per word and total pause duration. In other words, participants who scored higher on these temporal and cognitive effort indicators also rated the tasks as more physically demanding. These results are consistent with previous studies on the ergonomics of translation technologies, which suggest that suboptimal CAT tool interfaces can increase cognitive and physical demands on translators by necessitating repetitive and unnecessary actions (O'Brien et al. 2017). Taking the rTAPs into account, many participants from all groups expressed negative opinions about the digital environment, citing the absence of frequently used shortcuts and the unfamiliar interface. Considering translation as a situated activity, where cognitive processes are inseparable from the physical and social environment (Risku 2002), it is clear that ergonomic and situational factors inevitably influence translators' performance. This is reflected in the present study, where longer task durations and more frequent pauses were associated with higher physical demand. In this context, physical demand should not be interpreted solely as a requirement for physically exhausting activities. Conversely, cognitively demanding tasks can also generate high perceptions of physical demand, as observed in this study. Within the experimental framework, this perception may have been amplified by the unfamiliar setup, including the software (Translog-II) and hardware (computer, keyboard and mouse), the physical location itself (not participants' usual workplace), the need to maintain long-term focus with only five-minute breaks between tasks and the restriction on using dictionaries or internet resources.

Conclusion

This study bridges a gap in the literature by linking effort indicators with expertise in MTPE. This is achieved by combining behavioural, subjective and product-based measures while focusing on MTPE in the legal text genre. The study also evaluates MTPE in the under-researched English-Turkish language pair. Our findings revealed that experienced translators exerted higher levels of effort in MTPE, which may be considered a valid indicator of expertise. This finding contradicts the results reported in the previous studies (see Dong & Chen 2025; Weng & Zheng 2024, for further discussion). Supported by self-reports of task demand, qualitative interviews and accuracy scores, these outcomes emphasise that translators are part of an inseparable ecosystem in which MT is playing an increasingly central role, with MTPE operations becoming essential for achieving high-quality results. However, the high effort rates associated with MTPE underscore the need for targeted training to promote effective and time-efficient MTPE strategies. Future research could replicate the experiment with a larger sample size and investigate how training interventions, tool design and ergonomic improvements could reduce effort while maintaining quality. Overall, our findings enhance our understanding of how expertise influences effort in MTPE, providing empirical evidence to inform theory and training in translation studies. This is particularly important in a world where MT is being used more and more by translators for professional translation purposes, helping them to overcome physical and cognitive limitations with the aid of technology.

Acknowledgement

This article is based on the doctoral dissertation of Minel Sayar Öztürk, supervised by Assoc. Prof. Dr. Alper Kumcu, to be submitted to Hacettepe University, Graduate School of Social Sciences, Department of Translation and Interpreting in English.

Compliance with Ethics Guidelines

The study was approved by the Social Sciences Ethics Committee of Hacettepe University (no. E-66777842-300-00003055757, September 6, 2023)

Disclosure Statement

No potential conflict of interest was reported by the author(s).

References

Ateşman, E. (1997). Türkçe'de okunabilirliğin ölçülmesi. *Dil Dergisi, 58*, 71–74.

Carl, M. (2012, May). Translog-II: A program for recording user activity data for empirical reading and writing research. In N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Eighth International Conference on Language Resources and Evaluation* (LREC'12) (pp. 4108-4112). European Language Resources Association.

Dong, D., & Chen, M.-L. (2025). Metacognitive strategies in translation: A comparative study of student and professional translators. *Humanities and Social Sciences Communications*, *12*(1), 890. https://doi.org/10.1057/s41599-025-05289-7

Dragsted, B. (2004). *Segmentation in translation and translation memory systems: An empirical investigation of cognitive segmentation and effects of integrating a TM system into the translation process* (Doctoral dissertation, Copenhagen Business School). Samfundslitteratur.

Dragsted, B., Hansen, I. G., & Sørensen, H. S. (2009). Experts exposed. In F. Alves & S. Göpferich (Eds.), *Methodology, Technology and Innovation in Translation Process Research* (pp. 293–317). Samfundslitteratur.

Ehrensberger-Dow, M., Albl-Mikasa, M., Andermatt, K., Hunziker Heeb, A., & Lehr, C. (2020). Cognitive load in processing ELF: Translators, interpreters, and other multilinguals. *Journal of English as a Lingua Franca*, *9*(2), 217–238. https://doi.org/10.1515/jelf-2020-2039

Eszenyi, R. (2025). Post-editing van machinevertaling: Is dat een vorm van vertaling of revisie? *Roczniki Humanistyczne*, *73*(10, Zeszyt specjalny), 15–33. https://doi.org/10.18290/rh257310.1s

Gieshoff, A. C., & Heeb, A. H. (2023). Cognitive load and cognitive effort: Probing the psychological reality of a conceptual difference. *Translation, Cognition & Behavior*, *6*(1), 3–28. https://doi.org/10.1075/tcb.00073.gie

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): results of empirical and theoretical research. In *Advances in psychology* (Vol. 52, pp. 139–183). Elsevier. https://doi.org/10.1016/S0166-4115(08)62386-9

Herbig, N., Pal, S., Krüger, A., & van Genabith, J. (2021). Multi-modal estimation of cognitive load in post-editing of machine translation. In O. Czulo, S. Hansen-Schirra, R. Rapp, & M. Bisiada (Eds.), *Translation, interpreting, cognition: The way out of the box* (pp. 1–32). Language Science Press. https://doi.org/10.5281/ZENODO.4544686

Hutchins, W. J. (2001). Machine translation and human translation: in competition or in complementation? *International Journal of Translation*, *13*(1–2), 5–20.

Ibrahim, S. N. (2025). On assessing the accuracy of Arabic-English translation by machine and human. In M. H. Al Aqad (Ed.), *Advances in computational intelligence and robotics* (pp. 119–146). IGI Global. https://doi.org/10.4018/979-8-3373-0060-3.ch005

Immonen, S. (2006). Translation as a writing process: Pauses in translation versus monolingual

text production. *Target. International Journal of Translation Studies*, *18*(2), 313–336. https://doi.org/10.1075/target.18.2.06imm

Jakobsen, A. L. (2003). Effects of think aloud on translation speed, revision, and segmentation. In F. Alves (Ed.), *Triangulating translation: Perspectives in process-oriented research* (pp. 69–95). John Benjamins Publishing Company.

Kincaid, J. P., Fishburne, R. P., Jr., Rogers, R. L., & Chissom, B. S. (1975). Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel (Research Branch Report 8-75). *Chief of Naval Technical Training, Naval Air Station Memphis*.

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.

Koponen, M., Salmi, L., & Nikulin, M. (2019). A product and process analysis of post-editor corrections on neural, statistical and rule-based machine translation output. *Machine Translation*, *33*(1), 61–90. https://doi.org/10.1007/s10590-019-09228-7

Koponen, M., Sulubacak, U., Vitikainen, K., & Tiedemann, J. (2020, November). MT for subtitling: User evaluation of post-editing productivity. In A. Martins, H. Moniz, S. Fumega, B. Martins, F. Batista, L. Coheur, C. Parra, I. Trancoso, M. Turchi, A. Bisazza, J. Moorkens, A. Guerberof, M. Nurminen, L. Marg, & M. L. Forcada (Eds.), *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation* (EAMT 2020) (pp. 115–124). European Association for Machine Translation.

Krings, H. P. (1986). *Was in den Köpfen von Übersetzern vorgeht*. Gunter Narr.

Krings, H. P. (2001). *Repairing texts: Empirical investigations of machine translation post-editing processes*. Kent State University Press.

Kumpulainen, M. (2015). On the operationalisation of 'pauses' in translation process research. *Translation & Interpreting*, *7*(1), 47–58.

Lacruz, I. (2017). Cognitive effort in translation, editing, and post-editing. In J. W. Schwieter & A. Ferreira (Eds.), *The handbook of translation and cognition* (pp. 386–401). Wiley.

Lacruz, I., Shreve, G. M., & Angelone, E. (2012, October 28). Average pause ratio as an indicator of cognitive effort in post-editing: A case study. In S. O'Brien, M. Simard, & L. Specia (Eds.), *Proceedings of the Workshop on Post-Editing Technology and Practice, AMTA 2012* (pp. 1-10). Association for Machine Translation in the Americas.

Lörscher, W. (1991). Thinking-aloud as a method for collecting data on translation processes. In S. Tirkkonen-Condit (Ed.), *Empirical research in translation and intercultural studies* (pp. 67–77). Gunter Narr.

Muñoz Martín, R. (2014). Situating translation expertise: a review with a sketch of a construct. In J. W. Schwieter & A. Ferreira (Eds.), *The development of translation competence: Theories and methodologies from psycholinguistics and cognitive science* (pp. 2–54). Cambridge Scholars Publishing.

Mutta, M. (2016). Pausal behavior in the writing processes of foreign and native language writers: The importance of defining the individual pause length. In S. Plane, C. Bazerman, F. Rondelli, C. Donahue, A. N. Applebee, C. Boré, P. Carlino, M. Marquilló Larruy, P. Rogers, & D. Russell (Eds.), *Recherches en écriture: Regards pluriels* (Recherches textuelles, 13, pp. 583–604). Centre de Recherche sur les Médiations (Crem), Université de Lorraine.

O'Brien, S. (2006). Pauses as indicators of cognitive effort in post-editing machine translation output. *Across Languages and Cultures*, *7*(1), 1–21. https://doi.org/10.1556/Acr.7.2006.1.1

O'Brien, S. (2011). Towards predicting post-editing productivity. *Machine Translation*, *25*(3), 197–

215. https://doi.org/10.1007/s10590-011-9096-7

O'Brien, S. (2022). How to deal with errors in machine translation: post-editing. In D. Kenny (Ed.), *Machine translation for everyone: Empowering users in the age of artificial intelligence* (pp. 105–120). Language Science Press.

O'Brien, S., Ehrensberger-Dow, M., Connolly, M., & Hasler, M. (2017). Irritating CAT Tool Features that Matter to Translators. *HERMES - Journal of Language and Communication in Business*, *56*, 145–162. https://doi.org/10.7146/hjlcb.v0i56.97229

Paradowska, U., Sycz-Opon, J., & Badziński, A. (2025). MT use by professional translators in Poland – survey results. *Perspectives*, 1–22. https://doi.org/10.1080/0907676X.2025.2502937

Persky, A. M., & Robinson, J. D. (2017). Moving from novice to expertise and its implications for instruction. *American Journal of Pharmaceutical Education*, *81*(9), 6065. https://doi.org/10.5688/ajpe6065

Popović, M., Lommel, A., Burchardt, A., Avramidis, E., & Uszkoreit, H. (2014, June). Relations between different types of post-editing operations, cognitive effort and temporal effort. In M. Cettolo, M. Federico, L. Specia, & A. Way (Eds.), *Proceedings of the 17th Annual Conference of the European Association for Machine Translation* (pp. 191–198). European Association for Machine Translation.

R Core Team. (2025). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org

Revelle, W. (2018). *psych: Procedures for personality and psychological research* (Version 1.8.12) [R package]. https://CRAN.R-project.org/package=psych

Rico Pérez, C. (2024). Re-thinking machine translation post-editing guidelines. *The Journal of Specialised Translation*, *41*, 26–47. https://doi.org/10.26034/cm.jostrans.2024.4696

Risku, H. (2002). Situatedness in translation studies. *Cognitive Systems Research*, *3*(3), 523–533. https://doi.org/10.1016/S1389-0417(02)00055-4

Robson, S. G., Tangen, J. M., & Searston, R. A. (2021). The effect of expertise, target usefulness and image structure on visual search. *Cognitive Research: Principles and Implications*, *6*(1), 16. https://doi.org/10.1186/s41235-021-00282-5

Ruokonen, M., & Koskinen, K. (2017). Dancing with technology: Translators' narratives on the dance of human and machinic agency in translation work. *The Translator*, *23*(3), 310–323. https://doi.org/10.1080/13556509.2017.1301846

Sánchez-Torrón, M., Ipek, E., & Raído, V. E. (2025). Creating terminological correspondence recognition tests with GPT-4: a case study in English-to-Turkish translations in the engineering domain. *International Journal of Artificial Intelligence in Education*. https://doi.org/10.1007/s40593-025-00465-x

Šarčević, S. (1997). *New approach to legal translation*. Kluwer Law International.

Sharma, S., Diwakar, M., Singh, P., Singh, V., Kadry, S., & Kim, J. (2023). Machine translation systems based on classical-statistical-deep-learning approaches. *Electronics*, *12*(7), 1716. https://doi.org/10.3390/electronics12071716

Weng, Y., & Zheng, B. (2024). Automaticity in translation: Effects of time pressure and work experience. *Meta*, *69*(2), 355–379. https://doi.org/10.7202/1118382ar

Yanmei, L., & Mingmei, D. (2014). A multiple case study of Chinese-English translation strategies. *Studies in Literature and Language*, *9*(3), 58–65. http://dx.doi.org/10.3968/6098