Measuring cognitive effort in post-editing: an eye-tracking study comparing professional and student translators

Ana María Rojo López
María Inmaculada Vicente López
Kristian Tangsgaard Hvelplund

revista.tradumàtica
tecnologies de la traducció

Ana María Rojo López
Universidad de Murcia;
anarojo@um.es;
ORCID: 0000-0003-4303-9047

María Inmaculada Vicente López
Universidad de Murcia;
minmaculada.vicente@um.es;
ORCID: 0000-0002-6169-0805

Kristian Tangsgaard Hvelplund
University of Copenhagen;
kristian.hvelplund@hum.ku.dk;
ORCID: 0000-0001-7593-2177

Abstract

This eye-tracking study compares the post-editing cognitive effort of 25 professionals and 27 students when post-editing NMT and SMT from English to Spanish. Results show no significant differences in post-editing time or fixation duration between groups or MT systems, but reveal reduced fixation duration with NMT for both groups.

Keywords: eye tracking; post-editing; neural machine translation; cognitive effort; post-editing time.

Resumen

Este estudio de seguimiento ocular compara el esfuerzo cognitivo de post-edición de 25 profesionales y 27 estudiantes al post-editar la traducción automática neuronal (NMT, por sus siglas en inglés) y la traducción automática estadística (SMT, por sus siglas en inglés) del inglés al español. Los resultados no muestran diferencias significativas en el tiempo de post-edición ni en la duración de las fijaciones entre los grupos o los sistemas de TA, pero revelan una reducción en la duración de las fijaciones con NMT para ambos grupos..

Palabras clave: seguimiento ocular; post-edición; traducción automática neuronal (TAN); esfuerzo cognitivo; tiempo de posedición .

Resum

Aquest estudi de seguiment ocular compara l'esforç cognitiu de post-edició de 25 professionals i 27 estudiants en post-editar la traducció automàtica neuronal (NMT, per les sigles en anglès) i la traducció automàtica estadística (SMT, per les sigles en anglès) de l'anglès al castellà. Els resultats no mostren diferències significatives en el temps de post-edició ni en la durada de les fixacions entre els grups o els sistemes de TA, però revelen una reducció en la durada de les fixacions amb NMT per a tots dos grups.

Paraules clau: seguiment ocular; post-edició; traducció automàtica neuronal (TAN); esforç cognitiu; temps de post-edició .

Rebuda: 2 de agost de 2023 | Acceptació: 19 de març de 2024 | Publicació avançada: 15 de desembre de 2024

Revista Tradumàtica. Número 22 (Desembre 2024)
Citació recomanada: ROJO LÓPEZ, ANA MARÍA; VICENTE LÓPEZ, MARÍA INMACULADA; TANGSGAARD HVELPLUND, KRISTIAN (2024). «Measuring cognitive effort in post-editing: an eye-tracking study comparing professional and student translators». Revista Tradumàtica. Tecnologies de la Traducció, 22, 112-135. <https://doi.org/10.5565/rev/tradumatica.418>

http://revistes.uab.cat/tradumatica
ISSN 1578-7559

1. Introduction

In recent decades, the translation industry has undergone a swift transformation propelled by globalisation, automation, cost reduction, and the imperative for expeditious turnarounds (Nunes Vieira, 2018). Language service providers (LSPs) have adeptly adjusted to the evolving translation market landscape through the integration of novel technologies into their production and management systems. This adaptation includes the use of computer-assisted translation (CAT) and machine translation (MT) tools, with the emergence of large language models (LLM) such as GPT-4 on the horizon (Hendy et al., 2023; Wang et al., 2023). A consequence of the assimilation of MT is the escalating demand for incorporating MT post-editing into CAT tools (Bundgaard, 2017; Castilho et al., 2018; Sakamoto, 2019).

Post-editing research has garnered considerable attention within Translation Studies. Extant studies explore aspects related to both the post-editing process and product, such as task time and resultant product quality (Alves et al., 2016; Guerberof Arenas, 2008, 2014; Koponen & Salmi, 2017; Läubli et al., 2019; O'Brien, 2011; Sánchez-Gijón et al., 2019; Teixeira & O'Brien, 2017), as well as problem-solving and research strategies (Carl et al., 2015; Daems, 2016; Nitzke, 2019; Witczak, 2021). Most studies on the post-editing process explore post-editing time in comparison with CAT and MT systems, but further research on the cognitive effort involved is needed to elucidate the processes involved in such a complex cognitive task (O'Brien, 2017).

Most work in Cognitive Translation and Interpreting Studies (CTIS) explores post-editing cognitive effort using statistical machine translation (SMT) (Koponen, 2016; Koponen et al., 2012; Moorkens, 2018; Nunes Vieira, 2015) — already an outdated MT architecture — but research using neural machine translation (NMT) is still scarce, especially for a widely-used language pair like English-Spanish. Previous research on NMT post-editing has primarily focused on productivity, measured by seconds per segment (Sánchez-Gijón et al., 2019), and quality enhancements (Castilho et al., 2018) in comparison to SMT. Nevertheless, there remains a paucity of eye-tracking studies investigating the cognitive effort required for post-editing NMT, especially within the English-Spanish language pair.

This paper introduces an eye-tracking study that aims to bridge this gap by comparing student and professional translators' eye movements and task time during an English-to-Spanish post-editing task using NMT and SMT. Even if, as already indicated, SMT is somehow outdated, comparing NMT with SMT allows us to increase experimental control and can be useful to highlight how advancements in MT technology have impacted translators' cognitive effort. But before introducing the study, we will provide an overview of extant work on post-editing cognitive effort.

2. Cognitive effort in post-editing

Post-editing refers to the process of refining machine-generated translations with the aim of enhancing output quality (ISO 18587, 2017). In the course of post-editing, translators need to adhere to stipulated guidelines delineating requisite edits and edits to be avoided. These guidelines ensure that the resultant target text attains a predetermined level of quality, typically specified by the client. The expected level of quality plays a key role in defining the type of post-editing interventions required. According to TAUS guidelines, light post-editing involves making only indispensable edits to enhance the comprehensibility of machine-translated output, even if it may retain stylistic and grammatical errors. In contrast, full post-editing requires achieving a quality level equivalent to that of human translation, thereby rendering the final product suitable for publication (Massardo et al., 2016).

Research on post-editing effort emerged as a prominent area of interest among CTIS scholars about a decade ago, building upon Krings' (2001) work, where he established three interconnected dimensions of effort: temporal, technical, and cognitive. Studies identifying the factors that contribute to the level of effort expended in post-editing have garnered significant attention. Several studies have identified factors associated with increased levels of temporal effort and technical effort, such as sentence length (Koponen et al., 2012; Tatsumi, 2009; Tatsumi & Roturier, 2010), the lack of use of controlled language (Temnikova, 2010), sentence structure (Aziz et al., 2012; Nunes Vieira, 2015), MT output errors (Koponen et al., 2012) and professional experience (Guerberof Arenas, 2014).

More recent post-editing work in CTIS moves on from Krings' different dimensions and focuses on the notion of cognitive effort as a construct more complex than the mere amalgamation of technical and temporal effort (Mellinger & Hanson, 2018). Post-editing effort is understood as the mental effort invested by a translator when revising and editing a text previously translated by a MT engine (Lacruz et al., 2014). The level of effort may vary depending on individual variables such as experience or motivation. The notion of cognitive effort is distinguished from that of cognitive load, which is associated with the traits of the pre-translated text which demand some effort from the translator (Lacruz et al., 2014). Studies on post-editing in CTIS use tools such as eye trackers and keyloggers to measure eye movements and keystrokes as indicators of cognitive effort.

By analysing translators' eye movements in real time, eye tracking provides valuable insights to better understand and quantify post-editing cognitive effort. Eye movements are interpreted as indicators of cognitive effort based on the eye-mind assumption by Just & Carpenter (1980: 330), which posits that there is a relationship between visual focus and cognitive focus. Eye-tracking metrics commonly employed in CTIS are fixation durations, fixation counts, gaze time, and pupil dilation (Alves et al., 2016; Carl et al., 2015; Daems, 2016; Koglin, 2015; Moorkens, 2018; Nunes Vieira, 2015; Teixeira, 2014). These metrics serve as essential tools to investigate and analyse the relationship between observable behavioural manifestations of eye movements and the underlying cognitive processes involved in post-editing tasks. Eye-tracking research on post-editing has

identified professional experience as a key contributing factor that can potentially influence cognitive effort in post-editing. For instance, Daems (2016) reports a clear influence of professional translation experience on several indicators, such as average fixation duration on the source text, number of fixations on the target text, adequacy error weight, and the use of external resources, which had an impact on post-editing speed and quality. Results from the study reveal that students spend more time on post-editing and have longer fixation durations and higher fixation counts than professional translators. While results do not reach statistical significance, they indicate that post-editing seems to be more cognitively demanding for translation students than for professionals when working with Dutch and English. More recently, Stasimioti and Sosoni (2021) report similar results for the English-Greek language pair, with translation students also showing a higher number of fixations than professional translators.

Keylogging studies on post-editing (e.g. Koponen et al., 2012) use indicators of cognitive effort provided by keyloggers (keystrokes, type of edits made by translators, total post-editing time), or automatic MT metrics (human-targeted translation edit rate or HTER). Post-editing time can be used as a proxy for cognitive effort, based on the assumption that challenging segments or certain MT errors may lead to longer editing times, which can be potentially related to cognitive effort. The work by Koponen et al. (2012) shows that post-editing time should be considered in the analysis of cognitively difficult errors. This raises a question about the potential correlation between post-editing time and cognitive effort measured by eye-tracking indicators. Post-editing time is also used to demonstrate the benefits of NMT over previous MT systems in terms of time (and, consequently, cost) reductions (Ragni & Nunes Vieira, 2022). Results generally show that NMT post-editing is faster than using CAT tools to translate fuzzy segments (Sánchez-Gijón et al., 2019; Witczak, 2021) or new segments (Daems, 2016; Jia et al., 2019; Läubli et al., 2019; Nitzke, 2019). Post-editing NMT is also faster than post-editing SMT (Castilho et al., 2018; Koponen et al., 2019; Stasimioti & Sosoni, 2019). However, none of these studies is conducted in the English-Spanish language pair. NMT is also proven to be more efficient than SMT in terms of MT quality (Koponen et al., 2019; Popović et al., 2016; Stasimioti et al., 2020; Toral & Sánchez-Cartagena, 2017).

NMT displays higher quality and productivity than SMT, but its cognitive benefits remain unclear. Most research on post-editing cognitive effort focuses on SMT, and comparisons with NMT are still scarce. Data from the few extant studies point to lower cognitive effort when post-editing NMT as compared with SMT, but results are not conclusive. For instance, the keylogging study by Jia et al. (2019) shows that post-editing phrase-based statistical machine translation (PBSMT) is cognitively more demanding than NMT post-editing for the English-Chinese language pair. But Toral et al. (2018) report no significant difference in the duration of pauses when post-editing NMT and PBSMT from English to Catalan. Interestingly, pauses are even longer in NMT as compared to PBSMT. To the authors' knowledge, Stasimioti & Sosoni (2019) is the only study using an eye tracker to compare cognitive and temporal effort in NMT and SMT post-editing. They measure average fixation count, mean fixation duration in milliseconds (ms), and total gaze time (in minutes) in a particular area of interest (AOI). Their results show higher

average fixation count and longer fixation duration in SMT than NMT. However, differences are very small and not statistically significant.

3. The study

Our study complements previous research by using eye tracking to explore the cognitive effort experienced by professional and student translators during NMT vs. SMT post-editing. SMT serves as an experimental control and also adds to previous research comparing NMT with other MT architectures, given the preliminary nature of extant results, which suggest just a partial consensus on reductions in post-editing cognitive effort (Ragni & Nunes Vieira, 2022). Moreover, English to Spanish is a rather under-researched language pair in the extant literature on eye-tracking studies exploring post-editing cognitive effort.

3.1 Aims and hypotheses

Our eye-tracking study compared student and professional translators' cognitive effort during an English-to-Spanish post-editing task using NMT and SMT. Cognitive effort was operationalised in terms of participants' eye movements and time employed in the task. Eye movements were measured by examining eye fixation duration (see section 3.3 below). Post-editing time was measured, through the Tobii Studio screen recording feature, as the total time spent on the post-editing task, including internet searches.

Three hypotheses were posed:

H1. Post-editing time will be shorter for professional translators than for students, regardless of the MT system used. Time will be shorter when post-editing NMT than SMT.

H2. Participants' fixation duration will be shorter for professional translators than for students, regardless of the MT system used. Duration will be shorter when post-editing NMT than SMT.

H3. Post-editing time will correlate positively with participants' eye fixation duration.

3.2 Participants

We initially recruited a total of 52 participants (25 professional translators and 27 translation students). Following data collection and the subsequent data quality assurance process (outlined in section 3.5), we ultimately retained 38 participants (73%) for further analysis: 18 professional translators and 20 translation students.

The 18 professional translators (five males and 13 females) had Spanish as their L1 and their ages ranged from 27 to 62 years (median 41). They had professional translation experience of between four and 38 years (median 16.5) and were recruited from professional associations and networks in Spain. They all also had MT post-editing experience, ranging from one to 10 years (median seven), and stated that they provided

post-editing as a regular service, along with translation and editing. Their participation was compensated with a voucher from a popular e-commerce platform.

The 20 translation students (four males and 16 females) also had Spanish as their L1 and their ages ranged from 20 to 21 years (median 21). They were enrolled in the fourth year of the Translation and Interpreting Undergraduate Degree programme taught at the University of Murcia (Spain). All these participants were familiar with Trados Studio from a semester-long Translation Technology module. However, none had prior experience of post-editing. To address this, they received specialised training in MT post-editing, which included a theoretical session on MT and post-editing guidelines, followed by a practical post-editing session in Trados Studio. This training was integrated into their coursework and assessed as part of the Translation Technology module.

They all consented to participate and were informed of the possibility of withdrawing from the study at any time. The study protocol was approved by the Research Ethics Committee of the University of Murcia.

3.3 Materials and tools

The task consisted of a full post-editing of a translation of a newspaper article from the British newspaper *The Guardian* into Spanish. A newspaper article was chosen to reduce potential sampling biases, since professional translators usually work with more specialised texts. This procedure is in line with previous eye-tracking studies (Daems, 2016; Koponen et al., 2012; Nunes Vieira, 2015).

Considering there is scant eye-tracking research on post-editing cognitive effort using a commercial CAT tool interface (Ragni & Nunes Vieira, 2022), we chose SDL Trados Studio 2017[1] for the post-editing of the article. Trados Studio was deemed valuable for emulating the working conditions of professional translators and collecting data in a non-invasive way. We used a translation memory with no matches in Trados Studio and the file was pre-translated with MT output.

The source text (ST) consisted of 191 words divided into seven segments (see Appendix 1). All participants (students and professionals) were distributed into two groups: group A post-edited the ST pre-translated with NMT (DeepL)[2] (220 words of MT output), and group B post-edited the ST pre-translated with SMT (PROMT)[3] (214 words of MT output).

The texts were not very long, the intention being to capitalise on the substantial amount of eye-tracking data and to reduce the risk of participant fatigue during task execution. The use of short experimental texts aligns with the practice of using source text materials containing fewer than 200 words in eye-tracking translation studies (Daems, 2016; Hvelplund, 2011; Koglin, 2015; Nitzke, 2019; O'Brien, 2009).

[1] https://www.trados.com/products/trados-studio/
[2] https://www.deepl.com/translator
[3] https://www.online-translator.com. PROMT was a SMT engine at the time the study was conducted.

Participants' eye movements were recorded using a Tobii T120 eye tracker,[4] which is a remote eye tracker integrated into a 17" computer monitor. The Tobii T120 eye tracker collects gaze samples at a rate of 120 Hz with an accuracy of 0.5 degrees (Tobii© Technology 2010). Tobii Studio v. 3.3.2. was used for collecting, processing, and exporting gaze data. Task duration was measured, through the Tobii Studio screen recording feature, as the overall time spent on the post-editing task (including internet searches).

3.4 Experimental procedure

Data were collected across multiple stages from November 2019 to November 2021 (interrupted by COVID-19 lockdown measures and travel restrictions). The study was conducted at two different locations: eye-tracking data from professional translators were gathered at Equus Traducciones, a LSP located in Granada (Spain), while data from translation students were collected at a dedicated eye-tracking setup at the University of Murcia.

For all participants, data collection took place in a room with blinded windows and a stable source of artificial light. A laptop on which Trados Studio 2017 and Tobii Studio software were installed was used. To minimise the risk of poor-quality gaze data, both the laptop and the Tobii eye tracker monitor were secured in position on a stable table surface, and participants sat in fixed chairs. The height of each chair was individually adjusted to establish an optimal distance and height in relation to the eye tracker.

Before recording, individual eye-tracking calibration was carried out using Tobii's five-point calibration grid. Participants were then asked to read a detailed post-editing brief displayed onscreen, which instructed them to perform a full post-editing of a newspaper article in accordance with TAUS guidelines (Massardo et al., 2016). They were not informed about the type of MT engine (SMT or NMT) used to produce the MT output. No time constraints were imposed on task completion, and participants were free to use internet resources at their discretion.

Following task completion, the eye-tracking recordings underwent manual segmentation in Tobii Studio. AOIs, namely *SourceText, TargetText*, and *Internet*, were demarcated to identify different regions of the screen for subsequent analysis.

3.5 Quality of eye-tracking data

Two parameters, Fixation Duration and Gaze Time on Screen (GTS), were used to assess the quality of the collected eye-tracking data (Hvelplund, 2014).

Fixation duration thresholds have traditionally ranged from 100 to 200 ms, with specific values employed in prior studies including 100 ms (Nunes Vieira, 2015), 180 ms (Sjørup, 2013) and 200 ms (Hvelplund, 2011). For the present investigation, fixations below 150 ms were deemed "unusually short" and consequently excluded.

[4] https://www.tobii.com/

The GTS percentage, provided by Tobii Studio, served as an additional metric to evaluate the quality of the eye-tracking data. A lower GTS score implies limited visual orientation toward the screen area by the participant or potential lapses in recording all eye movements by the eye tracker (Hvelplund, 2014: 217). Extending Hvelplund's (2011) GTS threshold calculation, post-editing tasks with GTS scores falling more than one standard deviation below the mean were omitted from the study. Within the professional translator group, three participants were excluded due to low GTS scores of 49%, 54%, and 67% (GTS threshold: 68%). Likewise, within the student group, two participants were excluded due to low GTS scores of 25% and 47% (GTS threshold: 61%).

As an additional measure to ensure data quality, eye tracker recordings underwent manual scrutiny to identify potential anomalies. Within the student group, five participants were excluded from the study due to non-compliance with task instructions or performing the post-editing task outside the Trados Studio interface. Consequently, data from 38 participants (73%), consisting of 18 professional translators and 20 translation students, were retained for inclusion in the study.

3.6 Statistical methods

3.6.1 Mann-Whitney U-test

For the analysis of post-editing time data, the Mann-Whitney U-test (Mann & Whitney, 1947) was employed. The Mann-Whitney U-test serves as a non-parametric alternative to Student's t-test and is recommended when data lack normal distribution and the sample population is below 20 (Mellinger & Hanson, 2017). Given the deviation from normal distribution observed in the Shapiro-Wilk test for the two datasets — post-editing time of translation students and professional translators — the Mann-Whitney U-test was deemed more appropriate (Saldanha & O'Brien, 2014).

3.6.2 Linear mixed-effects regression analysis

Inferential statistics were applied using linear mixed-effects regression (LMER) analysis to examine participants' fixation duration. LMER analysis, in comparison to factorial ANOVA designs, proves to be more robust as it accommodates both fixed and random effects. Random effects enable the estimation of individual effects and dependencies between observations, facilitating the determination of whether group differences are significant over variations among individual participants (Balling, 2008: 183). In the context of this study, an observation refers to the recording of an eye movement, specifically the fixation duration. Utilising a 120 Hz eye tracker, 120 fixations are recorded per second, resulting in 120 observations per second for each eye.

LMER proves particularly valuable for experimental techniques requiring the analysis of repeated measurement data from different subjects (Daems et al., 2016; Hvelplund, 2011, 2017; Lehr & Hvelplund, 2020; Mellinger & Hanson, 2018; Nitzke, 2019; Nunes Vieira, 2015; Schmaltz et al., 2016). Two LMER models were fitted into the R statistical

environment (R Core Team, 2023) to investigate the impact of two independent variables (*Group* and *MTOutput*) on one dependent variable (*FixationDuration*, in ms). Given the positively skewed nature of the *FixationDuration* data, a logarithmic transformation was applied (Balling & Hvelplund, 2015). The packages employed for estimating the mixed-effects models were lme4 (Bates et al., 2015) and lmerTest (Kuznetsova et al., 2017).

3.6.3 Kendall rank correlation coefficient analysis

In accordance with the recommendation by Mellinger & Hanson (2017), Kendall's τ was employed in the correlational analysis, given its demonstrated superiority over other non-parametric tests. The Kendall's correlation test was executed in the R statistical environment to investigate the potential relationship between post-editing time and fixation duration.

3.7 Analysis of results

3.7.1 Post-editing time

Measurement of post-editing time involved calculating the total duration of performance of the task (in seconds), inclusive of the time specifically spent on post-editing and on internet searches. Our first hypothesis predicted that the task time would be comparatively shorter for professional translators than for students, regardless of the MT system used. Table 1 shows the results from descriptive statistics, featuring means and standard deviations:

		NMT		SMT	
		M	SD	M	SD
Professionals	Total time	1643	955	2034	1412
	Post-editing time	478	1164	728	1463
	Internet search time	479	507	571	748
Students	Total time	1695	946	2021	657
	Post-editing time	1086	584	1505	466
	Internet search time	609	682	516	298

Table 1. *Descriptive statistics: mean and standard deviation*

Table 2 displays the results of Mann-Whitney U-tests run to compare time between students and professional translators when post-editing NMT and SMT:

Students vs. professionals (NMT)			Students vs. professionals (SMT)		
	W	p		W	p
Total time	48	0.8383	Total time	57	0.3477
Post-editing time	39	0.6534	Post-editing time	53	0.5403
Internet search time	56	0.3913	Internet search time	57	0.3477

Table 2. *Total task time, post-editing time and time spent on internet searches for student vs. professional translators*

As the data in Table 2 show, no statistically significant difference in time was observed between students and professional translators. Even if this result refutes our hypothesis, it aligns with findings from Nitzke (2019), who also reported no significant differences in time between translation students and professional translators.

When comparing the two MT engines, no statistically significant difference in time was reported either. Table 3 displays the results of the Mann-Whitney U-test comparing time between NMT and SMT:

NMT vs. SMT (professionals)			NMT vs. SMT (students)		
	W	p		W	p
Total time	28	0.2893	Total time	35	0.2730
Post-editing time	32	0.4799	Post-editing time	28	0.1041
Internet search time	39	0.9296	Internet search time	43	0.6232

Table 3. *Total task time, post-editing time and time spent on internet searches for NMT vs. SMT*

Despite the lack of statistically significant differences, the averages in Table 1 suggest a tendency toward reduced time requirements when working with NMT as opposed to SMT. This pattern is observable for both the student and professional translator groups, with an average discrepancy in total task time of 391 seconds (almost seven minutes) for professional translators (Table 1. M SMT 2034 – M NMT 1643) and 326 seconds (five minutes) for students (Table 1. M SMT 2021 – M NMT 1695).

It is also possible to observe tendencies in the averages when comparing the time specifically devoted to post-editing or to internet searches. Concerning post-editing time, the average measurements indicate that both professional and student translators spent more time on post-editing SMT compared to NMT, with an average discrepancy of 250 seconds (approximately four minutes) for professionals (Table 1. M SMT 728 – M NMT 478) and 419 seconds (approximately seven minutes) for students (Table 1. M SMT 1505 – M NMT 1086). Moreover, students devoted more time to post-editing (both NMT and

SMT) than professional translators. As already outlined, none of these observed differences attained statistical significance.

Focusing on the time devoted to internet searches, the professional translators spent more time on the internet when post-editing SMT in comparison to NMT, with an average difference of 90 seconds (Table 1. M SMT 571 – M NMT 479). Translation students behaved differently: they spent more time on internet searches when post-editing NMT, with an average difference of 93 seconds (Table 1. M NMT 609 – M SMT 516). Professionals were faster and spent less time on the internet when post-editing NMT than when post-editing SMT. Students were also faster carrying out the whole task but spent more time on internet searches. This finding aligns with results obtained by Olalla-Soler (2018). Figure 1 provides a visual representation of the time percentage allocated to each task.

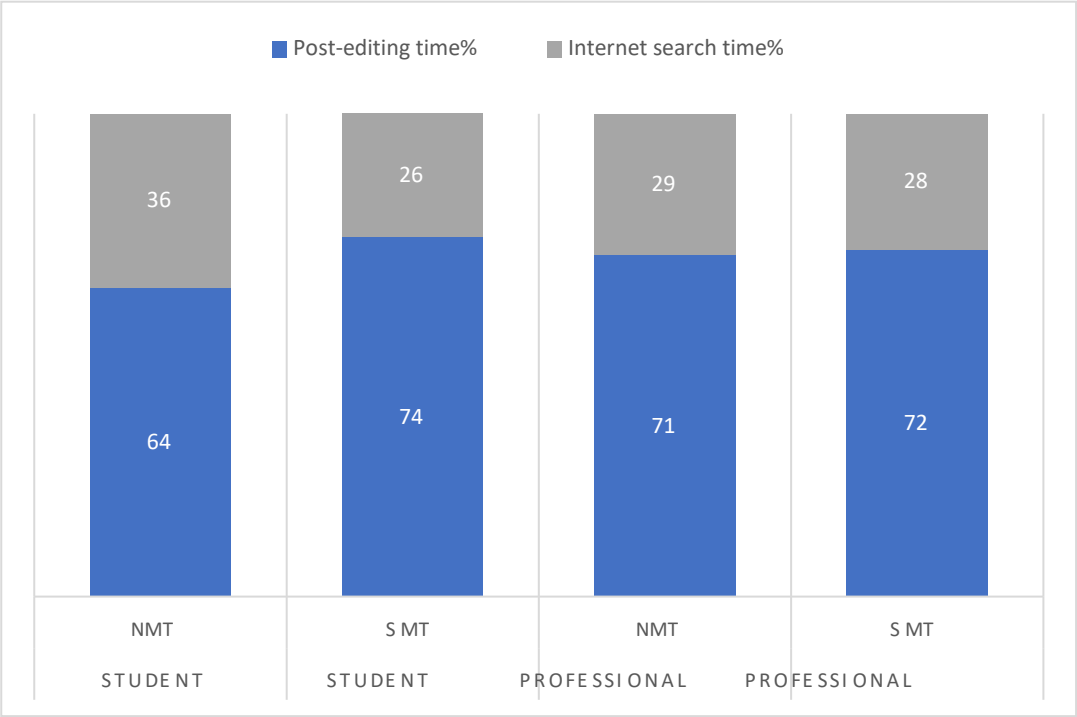

Figure 1. *Percentage of time devoted to post-editing and internet searches*

For professional translators, no noticeable difference was observed in the proportion of time devoted to post-editing or internet searches between NMT and SMT. Working with NMT did not result in changes to the post-editing behaviour of professional translators. In contrast, translation students invested more time in internet searches when post-editing NMT (36%) than when post-editing SMT (26%). The comparison between the two engines reveals greater differences between student and professional translators when post-editing NMT than when post-editing SMT.

3.7.2 Fixation duration

Our second hypothesis predicted that fixation duration would be lower for professional translators than for students, regardless of the MT system used. Table 4 displays descriptive statistics for fixation duration (in ms).

Group	NMT		SMT	
	M	SD	M	SD
Professionals	307	225	306	252
Students	285	214	299	211

Table 4. *Mean fixation duration (ms) per group and MT engine*

Contrary to our hypothesis, our descriptive findings unveiled marginally longer mean fixation durations (in ms) for professional translators than for students when post-editing the output of both MT engines.

In the inferential analysis, two LMER models were applied to test fixation duration differences between the two groups: professional and student translators. We created a data subset excluding fixation durations on the Internet AOI and divided it into two subsets: Subset 1 for NMT and Subset 2 for SMT. Subset 1 comprised 17,112 observations from 19 participants, while Subset 2 included 24,641 observations from 19 participants. For both LMER models, *FixationDuration* served as the dependent variable, and *Group,* categorised into *Professional* and *Student*, served as the independent variable. *Participant* was incorporated as the random effect, and *Group* was included as the fixed effect for the LMER models. The outcomes of the LMER analysis are detailed in Table 5.

Fixation duration Students vs. professionals	Estimate	Std. error	t	p
NMT	–0.05576	0.04177	–1.335	0.201
SMT	–0.03044	0.03032	–1.004	0.329

Table 5. *LMER analysis of the effect of Group on fixation durations*

As shown in Table 5, the LMER analysis revealed no statistically significant differences in fixation duration between students and professional translators, pointing to similar cognitive effort for both groups.

Regarding differences between the two MT engines, our second hypothesis also predicted that fixation duration would be lower when post-editing NMT than SMT. The descriptive statistics in Table 4 show lower fixations for students but not for professional translators.

For the LMER analysis, a data set including participants' fixation durations on both the NMT and SMT AOIs was generated. This subset was then split into two distinct sets: Subset 3, consisting of data from translation students, and Subset 4, comprising data from professional translators. Subset 3 included 71,664 observations from 20 participants, while Subset 4 included 71,187 observations from 18 participants. In both LMER models, *FixationDuration* served as the dependent variable, and *MTOutput*, categorised into NMT and SMT, served as the independent variable. *Participant* was incorporated as a random effect, and *MTOutput* was included as a fixed effect for both LMER models. The results of the LMER analysis are displayed in Table 6.

Fixation duration NMT vs. SMT	Estimate	Std. error	t	p
Professionals	0.03621	0.07696	4.705	0.00001
Students	0.03922	0.00753	5.209	0.00001

Table 6. *LMER analysis of the effect of MTOutput on fixation durations*

The LMER analyses corroborated the predicted difference between the MT engines. The tests revealed that fixation durations were significantly longer when post-editing SMT as compared to NMT for both students and professional translators, which suggests that using NMT in post-editing required less cognitive effort than using SMT. The contrast between the descriptive statistics and the LMER results may stem from the LMER test considering not only mean differences but also the variability within and between groups. This analytical approach enables an assessment of whether group differences extend beyond individual participant variations (Balling, 2008), emphasising the importance of using inferential analysis to validate findings obtained from descriptive statistics.

3.7.3 Correlation between post-editing time and cognitive effort

Our third hypothesis predicted a positive correlation between post-editing time and fixation duration. Table 7 displays the results of Kendall's τ correlations between post-editing time and mean fixation duration for both MT engines (NMT and SMT) and groups (students and professional translators).

	NMT		SMT	
	Kendall's τ	p	Kendall's τ	p
Students	0.0568	0.0000	0.0427	0.0000

Professionals	0.0548	0.0000	0.0486	0.0000

Table 7. *Kendall's τ correlations between overall post-editing time and fixation duration*

The analysis of Kendall's τ correlations between overall post-editing time and fixation duration revealed a very weak — almost negligible — correlation across the two MT engines and translator groups. The total time that participants spent on the task was not related to the cognitive effort invested, which contradicts the result from the keylogging study conducted by Koponen et al. (2012) on STM post-editing. Despite Koponen et al. measuring post-editing time at the sentence level, their results implied that shorter editing times were linked to errors rated as cognitively easier and to lower HTER scores.

4. Discussion and conclusion

This study compared the cognitive effort invested by student and professional translators during an English-to-Spanish post-editing task using NMT and SMT. Fixation duration and task time were measured as indicators of cognitive effort. The study also explored the relationship between the two indicators.

Our results on post-editing time did not corroborate our first hypothesis, since no significant difference in time was found between professional and student translators or between the two MT engines. On average, professional translators spent the same amount of time on post-editing as translation students. Similar outcomes were found by Nitzke (2019), possibly indicating that professional translators may invest more time in the task to maximise quality. However, given the limitations of our sample in terms of participants' experience (1-10 years), further research would be needed on translators with greater post-editing experience. No statistically significant difference in time was found between post-editing NMT and SMT either. A closer look at the time spent on internet searches provided insight into the matter. Even if professional translators spent less time on NMT, both post-editing and searching the internet, students still spent longer on internet searches when working with NMT than with SMT. The increased time allocated to NMT searches might be attributed to task complexity. Higher translation quality in NMT may have potentially allowed participants additional time to carefully review their revisions and intuitive decisions.

In our study, both students and professional translators allocated a substantial portion of their time to internet searches (accounting for approximately 28% to 29% in the case of professional translators and 26% to 36% in that of students), which matches the results observed by Nitzke (2019). Motivation for the use of external resources in post-editing is an issue that deserves further investigation. Despite the evidence that use of digital resources constitutes a significant portion of the total time spent on translation tasks (Hvelplund, 2017), there has been limited research on the use of external references during the post-editing process. While studies by Daems et al. (2016), Nitzke (2019), and Witczak (2021) initiated research in this direction, they employed different variables and

tasks, making their results non-comparable. Our findings raise the question of whether the quality of the MT output influences the number of internet searches during post-editing. Future investigations may explore whether lower-quality MT output results in more resource-intensive post-editing, while higher-quality output may lead to fewer internet searches.

Our results on fixation duration did not reveal statistically significant differences between students and professional translators. Nevertheless, our second hypothesis was partially confirmed, since fixation duration was significantly lower when working with NMT than with SMT, pointing to lesser investment of cognitive effort during the former. Potential explanations for the lack of differences between students and professional translators can be found in the working languages or the task difficulty. In contrast to prior research involving English and other languages, such as Dutch (Daems, 2016) or Greek (Stasimioti & Sosoni, 2021), where higher levels of cognitive effort were reported in students compared to professional translators, our findings revealed no statistically significant difference. It is worth noting that the specific language pair employed may influence cognitive effort outcomes. It is important to consider that the specific language pair used can impact cognitive effort. This is especially true for inflected languages like Spanish, where changing one word might require adjustments to verb forms or the cases of nouns and adjectives in a sentence. Even when post-editing high-quality machine translation output, such adjustments can significantly increase cognitive effort.

Another plausible explanation for the similar levels of cognitive effort reported in our study can be found in the task. Our professional translators reported having no previous experience translating or post-editing newspaper articles; in fact, only three of them had experience with similar text types, such as newsletters and blog postings. As suggested by Muñoz Martín (2014), specialised professional translators may find operating outside their narrow domain challenging, potentially explaining the similar cognitive effort levels observed. Translation experience is related to increased automaticity of translation processes, which is, in turn, associated with reduced cognitive effort (Hvelplund, 2016). Thus, our professional translators' limited experience with the text type may have prevented automatisation, contributing to their cognitive effort investment being comparable to that of students. Finally, our findings revealed a very weak correlation between fixation duration and task time. This contrasts with prior keylogging studies, such as Koponen et al. (2012) on post-editing or Jiménez-Crespo and Casillas (2021) on reviewing, where time and pauses were used as proxies for cognitive effort. Our results suggest that the relationship between post-editing time and cognitive effort may not be straightforward. Factors such as individual differences in post-editing behaviour, internet searches, or the language pair might play a role. While post-editing time is informative in gauging the challenge posed by machine translation output, it may not precisely reflect the actual cognitive effort involved. This raises questions about whether post-editing rates, based on the time spent on the task, align with post-editing cognitive effort in the same way as eye movements. Future research is needed to explore how compensation models based on time and technical effort, such as those using HTER, correspond to

actual cognitive effort in post-editing, considering that they only evaluate the final post-edited product and not the process (Ragni & Nunes Vieira, 2021).

Our research is not without challenges and limitations. One challenge was the difficulty of defining the post-editing experience of professional translators. Participants indicated their post-editing experience in years, consistent with previous studies on post-editing (Moorkens et al., 2015; Stasimioti & Sosoni, 2021). However, this does not fully prevent differences in post-editing experience, since translators may allocate different proportions of their time and work to different tasks. A limitation of our study is the relatively small number of participants in each group, i.e. 18 professionals and 20 students. Eye-tracking studies on the translation process often use small sample populations, ranging from around 25 participants (Hvelplund, 2011) to four or five (O'Brien, 2007; Moorkens, 2018). This tendency is primarily due to the inherent challenge of managing and analysing the extensive amount of eye movement data associated with eye-tracking technology. Consequently, the total of 38 participants in our study is notably high for this type of research. Another constraint pertains to the MT engines tested, with SMT being considered outdated. However, in the current study, SMT serves as a valuable benchmark for comparing results with NMT and plays a crucial role in illustrating the influence of technology on the cognitive effort of translators.

For the translation industry and CTIS research, the potential of emerging technologies, such as LLMs or generative pre-trained transformer (GPT) models, is undeniable. But, despite the limitations outlined and looking beyond the types of engines used, our study provides nuanced insights into differences in post-editing cognitive effort between students and professional translators, and proposes a solid methodological framework for implementing eye-tracking technology in a realistic setting, leveraging a CAT tool.

References

Alves, F., Koglin, A., Mesa-Lao, B., Martínez, M. G., de Lima Fonseca, N. B., de Melo Sá, A., Gonçalves, J. L., Szpak, K. S., Sekino, K., & Aquino, M. (2016). Analysing the impact of interactive machine translation on post-editing effort. In M. Carl, S. Bangalore, & M. Schaeffer (Eds.), *New Directions in Empirical Translation Process Research: Exploring the CRITT TPR-DB* (pp. 77–94). Springer Cham. <https://doi.org/10.1007/978-3-319-20358-4_4>. [Accessed: 20241212]

Aziz, W., Castilho, S., & Specia, L. (2012). PET: A tool for post-editing and assessing machine translation. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 3982–3987. <http://www.lrec-conf.org/proceedings/lrec2012/pdf/985_Paper.pdf>. [Accessed: 20241212]

Balling, L. W. (2008). A brief introduction to regression designs and mixed-effects modelling by a recent convert. *Copenhagen Studies in Language, 36*, 175–192. <https://doi.org/10.1075/ts.21013.sil>. [Accessed: 20241212]

Balling, L. W., & Hvelplund, K. T. (2015). Design and statistics in quantitative translation (Process) research. *Translation Spaces, 4*(1), 170–187. <https://doi.org/10.1075/ts.4.1.08bal>. [Accessed: 20241212]

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1). <https://doi.org/10.18637/jss.v067.i01>. [Accessed: 20241212]

Bundgaard, K. (2017). (Post-)Editing - A workplace study of translator-computer interaction at TextMinded Danmark A/S. [Doctoral Thesis, Aarhus University].

Carl, M., Gutermuth, S., & Hansen-Schirra, S. (2015). Post-editing machine translation: Efficiency, strategies, and revision processes in professional translation settings. In A. Ferreira & J. W. Schwieter (Eds.), *Psycholinguistic and Cognitive Inquiries into Translation and Interpreting* (pp. 145–174). John Benjamins Publishing Company. <https://doi.org/10.1075/btl.115.07car>. [Accessed: 20241212]

Castilho, S., Moorkens, J., Gaspari, F., Sennrich, R., Way, A., & Georgakopoulou, P. (2018). Evaluating MT for massive open online courses. *Machine Translation, 32*(3), 255–278. <https://doi.org/10.1007/s10590-018-9221-y>. [Accessed: 20241212]

Daems, J. (2016). A translation robot for each translator? A comparative study of manual translation and post-editing of machine translations: Process, quality and translator attitude. [Doctoral Thesis, Ghent University].

Daems, J., Vandepitte, S., Carl, M., & Jartsuiker, R. J. (2016). The effectiveness of consulting external resources during translation and postediting of general text types. In M. Carl, S. Bangalore, & M. Schaeffer (Eds.), *New Directions in Empirical Translation Process Research: Exploring the CRITT TPR-DB* (pp. 111–133). Springer Cham. <https://doi.org/10.1007/978-3-319-20358-4_6>. [Accessed: 20241212]

Guerberof Arenas, A. (2008). Productivity and quality in the post-editing of outputs from translation memories and machine translation. *Localisation Focus - The International Journal of Localisation, 7*(1), 11–21.

Guerberof Arenas, A. (2014). The role of professional experience in post-editing from a quality and productivity perspective. In S. O'Brien (Ed.), *Post-editing of Machine Translation: Processes and Applications* (pp. 51–76). Cambridge Scholars Publishing.

Hendy, A., Abdelrehim, M., Sharaf, A., Raunak, V., Gabr, M., Matsushita, H., Kim, Y. J., Afify, M., & Awadalla, H. H. (2023). How good are GPT models at machine translation? A comprehensive evaluation. <http://arxiv.org/abs/2302.09210>. [Accessed: 20241212]

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & van de Weijer, J. (2011). *Eye tracking: A comprehensive guide to methods and measures* (1st ed.). Oxford University Press.

Hvelplund, K. T. (2011). Allocation of cognitive resources in translation: An eye-tracking and key-logging study. [Doctoral Thesis, Copenhagen Business School].

Hvelplund, K. T. (2014). Eye tracking and the translation process: Reflections on the analysis and interpretation of eye-tracking data. In R. Muñoz Martín (Ed.), *MonTI. Monografías de Traducción e Interpretación* (Special Issue, pp. 201–223). Publicaciones de la Universidad de Alicante. <https://doi.org/10.6035/monti.2014.ne1.6>. [Accessed: 20241212]

Hvelplund, K. T. (2016). Cognitive efficiency in translation. In R. Muñoz Martín (Ed.), *Reembedding Translation Process Research* (pp. 149–170). John Benjamins Publishing Company. <https://doi.org/10.1075/btl.128.08hve>. [Accessed: 20241212]

Hvelplund, K. T. (2017). Translators' use of digital resources during translation. *Hermes (Denmark), 56*, 71–87. <https://doi.org/10.7146/hjlcb.v0i56.97205>. [Accessed: 20241212]

ISO 18587. (2017). Translation services — Post-editing of machine translation output — Requirements. Geneva: International Organization for Standardization. Retrieved from <https://www.iso.org/obp/ui/en/#iso:std:iso:18587:ed-1:v1:en>. [Accessed: 20241212]

Jia, Y., Carl, M., & Wang, X. (2019). Post-editing neural machine translation versus phrase-based machine translation for English–Chinese. *Machine Translation, 33*(1–2), 9–29. <https://doi.org/10.1007/s10590-019-09229-6>. [Accessed: 20241212]

Jiménez-Crespo, M. A., & Casillas, J. V. (2021). Literal is not always easier: Literal and default translation, cognitive effort, and comparable corpora. *Translation, Cognition and Behavior, 4*(1), 100–125. <https://doi.org/10.1075/tcb.00048.jim>. [Accessed: 20241212]

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review, 87*(4), 329–354. <https://doi.org/10.1037/0033-295X.87.4.329>. [Accessed: 20241212]

Koglin, A. (2015). An empirical investigation of cognitive effort required to post-edit machine translated metaphors compared to the translation of metaphors. *Translation & Interpreting, 7*(1).

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

Koponen, M., Aziz, W., Ramos, L., & Specia, L. (2012). Post-editing time as a measure of cognitive effort. *AMTA Workshop on Postediting Technology and Practice, 47*(3), 11–20.

Koponen, M., & Salmi, L. (2017). Post-editing quality: Analysing the correctness and necessity of post-editor corrections. *Linguistica Antverpiensia, New Series: Themes in Translation Studies, 16*(16), 137–148. <https://doi.org/10.52034/lanstts.v16i0.439>. [Accessed: 20241212]

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*. <https://doi.org/10.1007/s10590-019-09228-7>. [Accessed: 20241212]

Krings, H. P. (2001). *Repairing Texts: Empirical Investigations of Machine Translation Post-Editing Processes* (G. S. Koby, Ed.). The Kent State University Press.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in linear mixed effects models. *Journal of Statistical Software, 82*(13). <https://doi.org/10.18637/jss.v082.i13>. [Accessed: 20241212]

Lacruz, I., Denkowski, M., & Lavie, A. (2014). Cognitive demand and cognitive effort in post-editing. In S. O'Brien, M. Shimard & L. Specia (Eds.), *Proceedings of the 11th Conference of the Association for Machine Translation in the Americas* (pp. 73–84). Association for Machine Translation in the Americas.

Läubli, S., Amrhein, C., Düggelin, P., Gonzalez, B., Zwahlen, A., & Volk, M. (2019). Post-editing productivity with neural machine translation: An empirical assessment of speed and quality in the banking and finance domain. In *Proceedings of Machine Translation Summit XVII: Research Track* (pp. 267–272). <http://arxiv.org/abs/1906.01685>. [Accessed: 20241212]

Lehr, C., & Hvelplund, K. T. (2020). Emotional experts: Influences of emotion on the allocation of cognitive resources during translation. In R. Muñoz Martín & S. L. Halverson (Eds.), *Multilingual Mediated Communication and Cognition* (pp. 44–68). Routledge. <https://doi.org/10.4324/9780429323867-3>. [Accessed: 20241212]

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics, 18*(1), 50–60. <https://doi.org/10.1214/aoms/1177730491>. [Accessed: 20241212]

Massardo, I., van der Meer, J., O'Brien, S., Hollowood, F., Aranberri, N., & Drescher, K. (2016). MT postediting guidelines. TAUS Signature Editions.

Mellinger, C., & Hanson, T. (2017). *Quantitative Research Methods in Translation and Interpreting Studies*. Routledge. <https://doi.org/10.4324/9781315647845>. [Accessed: 20241212]

Mellinger, C. D., & Hanson, T. A. (2018). Order effects in the translation process. *Translation, Cognition & Behavior, 1*(1), 1–20. <https://doi.org/10.1075/tcb.00001.mel>. [Accessed: 20241212]

Moorkens, J. (2018). Eye tracking as a measure of cognitive effort for post-editing of machine translation. In C. Walker & F. M. Federici (Eds.), *Eye Tracking and Multidisciplinary Studies on Translation* (pp. 55–70). John Benjamins Publishing Company. <https://doi.org/10.1075/btl.143>. [Accessed: 20241212]

Moorkens, J., O'Brien, S., da Silva, I., Fonseca, N., & Alves, F. (2015). Correlations of perceived post-editing effort with measurements of actual effort. *Machine Translation, 29*, 267–284. <https://doi.org/10.1007/s10590-015-9175-2>. [Accessed: 20241212]

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. 2–56). Cambridge Scholars Publishing.

Nitzke, J. (2019). Problem-solving activities in post-editing and translation from scratch: A multi-method study. In *New Empirical Perspectives on Translation and Interpreting (Translatio)*. Language Science Press. <https://doi.org/10.5281/zenodo.2546446-5>. [Accessed: 20241212]

Nunes Vieira, L. (2015). Cognitive effort in post-editing of machine translation: Evidence from eye movements, subjective ratings, and think-aloud protocols. [Doctoral Thesis, Newcastle University].

Nunes Vieira, L. (2018). Automation anxiety and translators. *Translation Studies, 13*, 1–21. <https://doi.org/10.1080/14781700.2018.1543613>. [Accessed: 20241212]

O'Brien, S. (2007). Eye-tracking and translation memory matches. *Perspectives: Studies in Translatology, 14*(3), 185–205. <https://doi.org/10.1080/09076760708669037>. [Accessed: 20241212]

O'Brien, S. (2009). Eye tracking in translation-process research: methodological challenges and solutions. En I. M. Mees, S. Göpferich y F. Alves (Eds.), Methodology, Technology and Innovation in Translation Process Research. A Tribute to Arnt Lykke Jakobsen (Copenhagen). Samfundslitteratur Press. <https://doi.org/10.1075/target.24.1.13tir>. [Accessed: 20241212]

O'Brien, S. (2011). Towards predicting post-editing productivity. *Machine Translation, 25*, 197–215. <https://doi.org/10.1007/s10590-011-9096-7>. [Accessed: 20241212]

O'Brien, S. (2017). Machine translation and cognition. In J. W. Schwieter & A. Ferreira (Eds.), *The Handbook of Translation and Cognition* (1st ed., pp. 311–331). <https://doi.org/10.1002/9781119241485.ch17>. [Accessed: 20241212]

Olalla-Soler, C. (2018). Using electronic information resources to solve cultural translation problems: Differences between students and professional translators. *Journal of Documentation, 74*(6), 1293–1317. <https://doi.org/10.1108/JD-02-2018-0033>. [Accessed: 20241212]

Popović, M., Arčan, M., & Lommel, A. (2016). Potential and limits of using post-edits as reference translations for MT evaluation. *Proceedings of the 19th Annual Conference of the European Association for Machine Translation, EAMT 2016*, 218–229.

R Core Team. (2023). R: A language and environment for statistical computing. R Foundation for Statistical Computing. <https://www.R-project.org/>. [Accessed: 20241212]

Ragni, V., & Nunes Vieira, L. (2022). What has changed with neural machine translation? A critical review of human factors. *Perspectives, 30*(1), 137–158. <https://doi.org/10.1080/0907676X.2021.1889005>. [Accessed: 20241212]

Sakamoto, A. (2019). Why do many translators resist post-editing? A sociological analysis using Bourdieu's concepts. *The Journal of Specialised Translation, 31*, 201–216.

Saldanha, G., & O'Brien, S. (2014). Research methodologies in translation studies. In *Research Methodologies in Translation Studies*. <https://doi.org/10.4324/9781315760100>. [Accessed: 20241212]

Sánchez-Gijón, P., Moorkens, J., & Way, A. (2019). Post-editing neural machine translation versus translation memory segments. *Machine Translation, 33*(1–2), 31–59. <https://doi.org/10.1007/s10590-019-09232-x>. [Accessed: 20241212]

Schmaltz, M. S., da Silva, I. A. L., Pagano, A. S., Alves, F., Leal, A. L., Wong, D. F., Chao, L. S., & Quaresma, P. (2016). Cohesive relations in text comprehension and production: An exploratory study comparing translation and post-editing. In M. Carl, S. Bangalore, & M. Schaeffer (Eds.), *New Directions in Empirical Translation Process Research: Exploring the CRITT TPR-DB* (pp. 239–263). Springer Cham. <https://doi.org/10.1007/978-3-319-20358-4_11>. [Accessed: 20241212]

Sjørup, A. C. (2013). Cognitive effort in metaphor translation: An eye-tracking and key-logging study. Copenhagen Business School.

Stasimioti, M., & Sosoni, V. (2019). MT output and post-editing effort: Insights from a comparative analysis of SMT and NMT output for the English to Greek language pair and implications for the training of post-editors. In C. Szabó & R. Besznyák (Eds.), *Teaching Specialised Translation and Interpreting in a Digital Age - Fit-For-Market Technologies, Schemes and Initiatives* (pp. 151–175). Vernon Press.

Stasimioti, M., & Sosoni, V. (2021). Investigating post-editing: A mixed-methods study with experienced and novice translators in the English-Greek language pair. In Tra&Co Group (Ed.), *Translation, interpreting, cognition: The way out of the box* (pp. 79–104). Language Science Press. <https://doi.org/10.5281/zenodo.4545037>. [Accessed: 20241212]

Stasimioti, M., Sosoni, V., Mouratidis, D., & Kermanidis, K. (2020). Machine translation quality: A comparative evaluation of SMT, NMT and tailored-NMT outputs. *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation, EAMT 2020*, 441–450.

Tatsumi, M. (2009). Correlation between automatic evaluation metric scores, post-editing speed, and some other factors. *Proceedings of MT Summit XII, 2001*, 332–339. <http://www.mt-archive.info/MTS-2009-Tatsumi.pdf>. [Accessed: 20241212]

Tatsumi, M., & Roturier, J. (2010). Source text characteristics and technical and temporal post-editing effort: What is their relationship? *Second Joint EM+/CNGL Workshop "Bringing MT to the User: Research on Integrating MT in the Translation Industry" JEC 2010*, 43–51. Retrieved from http://www.mt-archive.info/JEC-2010-Tatsumi.pdf

Teixeira, C. S. C. (2014). The impact of metadata on translator performance: How translators work with translation memories and machine translation. [Doctoral Thesis, Universitat Rovira i Virgili]. <https://doi.org/10.13140/RG.2.1.2190.2887>. [Accessed: 20241212]

Teixeira, C. S. C., & O'Brien, S. (2017). Investigating the cognitive ergonomic aspects of translation tools in a workplace setting. *Translation Spaces, 6*(1), 79–103. <https://doi.org/10.1075/ts.6.1.05tei>. [Accessed: 20241212]

Temnikova, I. (2010). Cognitive evaluation approach for a controlled language post-editing experiment. *Proceedings of the 7th International Conference on Language Resources and Evaluation, LREC 2010*, January 2010, 3485–3490.

Toral, A., & Sánchez-Cartagena, V. M. (2017). A multifaceted evaluation of neural versus phrase-based machine translation for 9 language directions. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, 1063–1073. <https://doi.org/10.48550/arXiv.1701.02901>. [Accessed: 20241212]

Toral, A., Wieling, M., & Way, A. (2018). Post-editing effort of a novel with statistical and neural machine translation. *Frontiers in Digital Humanities, 5*. <https://doi.org/10.3389/fdigh.2018.00009>. [Accessed: 20241212]

Wang, L., Lyu, C., Ji, T., Zhang, Z., Yu, D., Shi, S., & Tu, Z. (2023). Document-level machine translation with large language models. ArXiv, abs/2304.02210. <https://api.semanticscholar.org/CorpusID:257952312>. [Accessed: 20241212]

Witczak, O. (2021). Information searching in the post-editing and translation process [Doctoral Thesis, Adam Mickiewicz University]. <https://doi.org/10.13140/RG.2.2.15691.87847>. [Accessed: 20241212]

APPENDIX 1. Comparative table of source text and MT output

#	Source text 191 words	NMT 220 words	SMT 214 words
1	Cigarette plain packaging laws come into force in Australia	Entran en vigor en Australia las leyes de empaquetado sencillo de cigarrillos	La llanura del cigarrillo leyes de embalaje entra en vigor en Australia

2	Australia's world-first laws on cigarette and tobacco plain packaging have come into force, replacing brand logos and colours with generic drab olive green coverings, gruesome pictures of diseased body parts and depictions of children and babies made ill by their parents' smoking.	Han entrado en vigor las primeras leyes australianas sobre el empaquetado de cigarrillos y tabaco, que sustituyen los logotipos y colores de las marcas por recubrimientos genéricos de color verde oliva monótono, imágenes horripilantes de partes del cuerpo enfermas y representaciones de niños y bebés enfermos a causa del tabaquismo de sus padres.	Las primeras leyes mundiales de Australia de cigarrillo y embalaje de la llanura de tabaco han entrado en vigor, sustituyendo logotipos de la marca y colores con cubiertas verdes verdes oliva apagadas genéricas, cuadros espantosos de partes del cuerpo enfermas y pinturas de niños y los bebés hicieron malo por sus padres fumar.
3	Apart from the varying health warnings and images the only difference between the packs, mandatory from Saturday, are the brand names, and these are all printed in identical small font.	Aparte de las diferentes advertencias e imágenes sanitarias, la única diferencia entre los envases, obligatorios a partir del sábado, son las marcas, y todas ellas están impresas en idéntico tipo de letra pequeña.	Aparte de las advertencias de la salud variadas e imágenes la única diferencia entre los paquetes, obligatorios a partir del sábado, es las marcas registradas, y éstos todos se imprimen en la pequeña fuente idéntica.
4	It is the world's most strict regime for the packaging of tobacco.	Es el régimen más estricto del mundo para el envasado de tabaco.	Es el régimen del mundo más estricto para el embalaje de tabaco.
5	Australia's federal government says the aim is to deter young people from smoking by stripping the habit of glamour.	El gobierno federal de Australia dice que el objetivo es disuadir a los jóvenes de fumar despojándolos del hábito del glamour.	El gobierno federal de Australia dice que el objetivo es desalentar a jóvenes de fumar por despojar el hábito del encanto.
6	It is relying on studies showing that if people have not started smoking by age 26 there is a 99% chance they will never take it up.	Se basa en estudios que muestran que si las personas no han comenzado a fumar a los 26 años de edad hay un 99% de probabilidades de que nunca lo hagan.	Confía en estudios mostrando que si la gente no ha comenzado a fumar por la edad 26 hay una posibilidad del 99% nunca lo tomarán.

7	Even from a very early age you can see that kids understand the message that the tobacco company is trying to sell through their branding, said the federal health minister, Tanya Plibersek, citing studies that showed, for example, children linking a crown in a logo with the idea of being a princess.	Incluso desde una edad muy temprana se puede ver que los niños entienden el mensaje que la compañía tabacalera está tratando de vender a través de su marca, dijo la ministra federal de salud, Tanya Plibersek, citando estudios que mostraban, por ejemplo, que los niños vinculan una corona en un logotipo con la idea de ser una princesa.	Incluso de una edad muy temprana puede ver que los niños entienden el mensaje que la empresa tabacalera trata de venderse a través de su marcación, dijo la ministra de salud federal, Tanya Plibersek, citando estudios que mostraron, por ejemplo, a niños que unen una corona en un logotipo con la idea de ser una princesa.