REVISTA DE LLENGUA I DRET // JOURNAL OF LANGUAGE AND LAW #78

POST-EDITING VS. TRANSLATING IN THE LEGAL CONTEXT: QUALITY AND TIME EFFECTS FROM ENGLISH TO SPANISH

Jeffrey Killman[*]

Mónica Rodríguez-Castro[**]

Abstract

Limited research has addressed the use of machine translation (MT) with legal texts, while recent data-driven MT approaches have improved quality. The present study reports on the results of an experiment involving 26 translators who post-edited and translated legal texts from English to Spanish. The results show that post-editing led to quality and time gains. The quality gains were statistically significant. Quality was determined using standardized scoring criteria from a professional association, and the time taken to complete the tasks was logged using Translog-II. Participants completed questionnaires before and after the experiment to gather data about demographic variables, previous experience, perceptions of MT post-editing and their reflections on the experiment. There was general agreement among the participants that the machine translation output was beneficial, particularly at the terminological and phraseological levels. However, participant variables related to years of experience or translation credentials did not reveal a demonstrable impact on quality and time.

Keywords: machine translation; neural machine translation; statistical machine translation; legal translation; post-editing.

POSTEDITAR O TRADUIR EN EL CONTEXT JURÍDIC: EFECTES SOBRE LA QUALITAT I EL TEMPS DE L'ANGLÈS A L'ESPANYOL

Resum

La recerca sobre l'ús de la traducció automàtica (TA) amb textos jurídics és escassa, tot i que els sistemes de TA més recents basats en dades han millorat la qualitat de les traduccions. En aquest estudi es presenten els resultats d'un experiment en què van participar 26 professionals de la traducció que van posteditar i traduir textos jurídics de l'anglès a l'espanyol. Els resultats mostren que la postedició va fer guanyar qualitat i temps. La millora de la qualitat era estadísticament significativa. La qualitat es va determinar mitjançant els criteris de puntuació estandarditzats d'una associació professional i el temps necessari per acabar les tasques es va registrar amb Translog-II. Les persones participants van contestar uns qüestionaris abans i després de l'experiment per recopilar dades sobre les variables demogràfiques, l'experiència prèvia, les opinions sobre la postedició de TA i les seves reflexions sobre l'experiment. Entre les respostes hi va haver un consens ampli sobre el fet que els resultats de la traducció automàtica eren positius, sobretot pel que fa a la terminologia i la fraseologia. Tanmateix, les variables dels participants relatives als anys d'experiència o les competències en traducció no van mostrar una incidència significativa ni en la qualitat ni en el temps.

Paraules clau: traducció automàtica; traducció automàtica neuronal; traducció automàtica estadística; traducció jurídica; postedició.

[*] Jeffrey Killman, associate professor of Spanish at the University of North Carolina at Charlotte, jkillman@uncc.edu, 0000-0002-5517-4118

[**] Mónica Rodríguez-Castro, associate professor of Translation Studies and Spanish at the University of North Carolina at Charlotte, monica.rodriguez@uncc.edu, 0000-0003-4067-0693

Article received: 10.05.2022. Blind reviews: 25.05.2022 and 22.08.2022. Final version accepted: 12.09.2022.

Recommended citation: Killman, Jeffrey, & Rodríguez-Castro, Mónica. (2022). Post-editing vs. translating in the legal context: Quality and time effects from English to Spanish. *Revista de Llengua i Dret, Journal of Language and Law*, *78*, 56-72, https://www.doi.org/10.2436/rld.i78.2022.3831

Generalitat de Catalunya
Escola d'Administració Pública
de Catalunya

Contents

1 Introduction

Data-driven approaches have contributed to significant quality gains in the output of machine translation (MT) systems in the past couple of decades. In the 2000s, the first data-driven paradigm shift was marked by statistical MT (SMT) (Koehn et al., 2003), whereas the state of the art before then had been rule-based MT (RBMT). MT in this new paradigm would no longer rely on grammatical rules or lexicographical or dictionary-derived input. Instead, primary data sources would produce statistically driven translations by drawing on large amounts of corpus data comprising millions of equivalent sentences aligned in different languages. The most recent paradigm shift, neural MT (NMT) is also data-driven "but uses a completely different computational approach" (Forcada, 2017, p. 292). NMT employs a more syntagmatic, multilayered approach that attempts to fill potential contextual gaps left by the more paradigmatic SMT approach that is more limited to the accuracy of smaller chunks of text that it strings together. Either way, these considerable data-driven strides have reinvigorated interest in how translations might be produced with different levels of MT and human intervention in a range of professional settings, content and/or domain areas.

Both SMT and NMT rely on statistics according to which a target sentence has a certain probability of being a translation of a source sentence (Forcada, 2017, p. 300). In the case of SMT, chunking consists of phrases that do not necessarily need to be categorizable as grammatical units (Koehn, 2010, p. 127; Kenny & Doherty, 2014, p. 284; Forcada, 2017, p. 301). The translations of these phrases are linked together in the output. First, the source words are aligned to the target words of these phrases according to probabilities determined from the bilingual corpus, then source and target phrases compatible with these word alignments are identified and assigned scores (Forcada, 2017, p. 301). The scores from this tuning process (Kenny & Doherty, 2014, p. 283) are complemented with additional probabilities derived from large corpora in the target language to help select probable target phrases. Longer phrases, which may include more relevant context, are prioritized, but they are not as frequent and hence not as statistically reliable (Koehn, 2010, p. 141). There is a risk they might not be selected by the system even if they exist.

For its part, NMT relies on neural networks "composed of thousands of artificial units that resemble neurons in that their output or activation […] depends on the stimuli they receive from other neurons and the strength of the connections along which these stimuli are passed" (Forcada, 2017, p. 292). Activations of individual neural units are combined with the activation of other neural units in layers consisting of hundreds of neural units, and layers are connected to other layers by weights allowing for connections to range in the thousands (Forcada, 2017, p. 295). These activation states "are trained to build distributed representations of words and their contexts, both in the context of the source sentence being processed and in the context of the target sentence being produced" (Forcada, 2017, pp. 293–294). These representations are more multidimensional than in the case of SMT, with translation being approached from a more holistic standpoint that is not primarily limited to immediately surrounding co-text. While NMT may be a deeper approach, it may risk probing too far into different dimensions and thus distance itself from potentially more related contexts.

Regardless of MT architecture, the probability that a target sentence is a translation of a given source sentence is potentially augmented the more related the corpora from which the systems draw information. Systems can operate optimally when there are sufficiently voluminous, directly related sources of corpora on which to adequately train them. Nevertheless, such systems are not always widely available or may not be developed outside specific institutions to assist with internal translation needs and workflows. In any event, a good amount of the data on which data-driven systems are trained comes from institutions such as the European Union or the United Nations (e.g., Koehn, 2005, 2010; Crego et al., 2016; Junczys-Downmunt, Dwojak, & Hoang, 2016; Koehn & Knowles, 2017), which produce large amounts of content relating to matters of law and justice[1]. Moreover, during initial development phases, the quality of domain-specific systems is often compared with that of commercially available or general-purpose systems, which speaks to the considerable level of quality one might expect from extensively trained, general-purpose systems in a variety of domains.

[1] The authors recognize that EU or UN data sources may exist only in these institutions' official languages or may be abundant in only a subset of these official languages. While it is possible to build MT systems in lower resourced languages, there are challenges involved in building them (Forcada, 2020).

In the legal translation context, more research has begun to address the impact of MT, data-driven or otherwise, on translator productivity and translation quality. A strand of the extant scholarship that addresses the topic focuses on MT quality issues in various legal translation contexts, with some studies weighing the value of MT in legal translation training (Heiss & Soffritti, 2018; Mileto, 2019; Wiesmann, 2019; Roiss, 2021) and others approaching quality from the standpoint of an MT user (Lewis, 1997; Yates, 2006; Kit & Wong, 2008; Killman, 2014; Dik, 2020) or that of a builder of systems (Gotti et al., 2008; Koehn & Knowles, 2017). In addition, a series of post-editing studies include domain-specific MT systems trained in specific legal contexts (Farzindar & Lapalme, 2009; Vardaro et al., 2019; Arnejšek & Unk, 2020; Macken et al., 2020; Stefaniak, 2020; Desmet, 2021) or include general-purpose MT and legal and other source texts (García, 2010, 2011; Şahin & Dungan, 2014). Finally, a few studies report on in-house translators' uptake of the domain-specific MT system implemented by the Directorate-General for Translation (DGT) (Cadwell et al., 2016; Lesznyák, 2019; Rossi & Chevrot, 2019).

The present study seeks to complement this growing area of studies by building on post-editing research involving legal texts. Similar to several of these studies, Google Translate (GT) was used when it was a statistical system. In this particular case, GT was used not long before it was replaced with a neural-net-based architecture, and the study involved the English–Spanish language pair, which has yet to be covered in a legal translation post-editing study. The legal texts were post-edited and translated by a set of participants, and the quality of the post-edited and translated products was systematically measured and compared, as was the time taken to complete the post-editing and translation tasks. Legal translation has long been regarded a particularly challenging area of specialized translation (e.g., Alcaraz & Hughes, 2002), and reliable translation-specific resources have not been historically plentiful in this area. Given that much of the corpora on which data-driven MT systems are trained originates from institutions that translate legal content in multiple language pairs including English–Spanish, the present study asks whether general-purpose data-driven MT at the height of its development during its phrase-based statistical phase might contribute to productivity. The current study attempts to answer this question with human participants in an experiment controlling for time and quality in a language pair forming part of GT's first stage of development.

2 Related research

An increasing body of research has specifically looked at the quality or potential use of MT with legal texts. In earlier studies, Lewis (1997), Yates (2006) and Kit and Wong (2008) provided insights on the performance of RBMT. Lewis, who investigated English-to-German and -French, and Yates, who examined German- and Spanish-to-English, both assessed a selection of output examples and pointed out a number of errors. Yates (2006), however, found the German-to-English output less deficient than the Spanish-to-English output from the system she used in her study: Babel Fish (provided by Systran). Lewis (1997) and Yates (2006) relied on their own human evaluations. In contrast, Kit and Wong (2008) followed up on such studies by applying popular automatic metrics (BLEU and NIST[2]) to several EU and UN documents available in a wide variety of language combinations that were compatible with the systems they evaluated (i.e., Babel Fish, GT, ProMT, SDL free translator, Systran and WordLingo). The goal in this study was to compare a range of engines in this area of legal translation from a variety of languages to English. Score averages were found to be highest in translations from French and Dutch. In translations from Spanish, scores were more middle-of-the-road. In general, the systems tended to perform relatively similarly in each language pair.

In the era of data-driven MT, most evaluative studies involving legal texts have come about in the recent wake of NMT (Koehn & Knowles, 2017; Heiss & Soffritti, 2018; Wiesmann, 2019; Dik, 2020; Roiss, 2021), while a few can be traced to the SMT paradigm (Gotti et al., 2008; Killman, 2014; Mileto, 2019). All but one of these NMT studies include output from DeepL, a general-purpose online system, in the Dutch–English, German–Italian, and German–Spanish language pairs (Heiss & Soffritti, 2018; Wiesmann, 2019; Dik, 2020; Roiss, 2021). Wiesmann also included MateCat: a workbench that can use output from DeepL, the NMT

[2] Bilingual evaluation understudy (BLEU) (Papineni et al., 2002), a widely used automatic metric for evaluating the quality of MT output, measures how much output overlaps with a reference translation using a score of 0–1. Based on BLEU, the National Institute of Standards and Technology (NIST) developed another such metric (Doddington, 2002). Kit and Wong (2008, p. 310) point out that their reasoning for using both metrics was because BLEU focuses more on fluency and NIST on lexical accuracy.

version of GT, and from Microsoft Translator, which was still an SMT system at the time. DeepL has been defined as having potential (Heiss & Soffritti, 2018; Dik, 2020) or is not regarded with optimism due to lexical, terminological or register shortcomings (Roiss, 2021, p. 503). Wiesmann (2019) found DeepL output to be superior in her study but reached the conclusion that MT is at a stage where post-editing effort would be quite substantial in the case of legal texts.

The study conducted by Koehn and Knowles (2017) stands out because it systematically compares NMT and SMT with automatic metrics (BLEU) in five areas including the legal domain. Koehn and Knowles compared the German-to-English quality of systems they trained with corpora from these five areas. Output was obtained from systems trained on corpora that were related (in-domain) and not related (out-of-domain) to the text being translated, which was subsampled from the corpora used to train each system. However, the subsampled tests were excluded from the training of the systems for test validity purposes. Though in-domain results were largely similar, SMT outperformed NMT in the legal, medical and religious (i.e., Quran) domains but not in the areas of IT and subtitles. Out-of-domain tests revealed NMT as inferior in nearly every case, and to a significant extent in some cases (Koehn & Knowles, 2017, pp. 29–30). In the tests where systems were trained on all five corpora, SMT performed better than NMT in the areas of law and IT, while the opposite was true of subtitles and the medical domain. In the case of the Quran, both systems were equal when trained on all five corpora. When it was trained on all five corpora, SMT outperformed the legal in-domain trained NMT system in the legal domain.

Gotti et al. (2008) illustrated early on how SMT accuracy in the legal domain could be increased when a system is trained on highly related corpora. Their system, TransLI (Translation of Legal Information), was designed to assist the Canadian federal courts in meeting the requirement to provide English and French translations of judgements. By training this system with data from the same courts, they were able to attain positive results according to various automatic metrics, revealing that TransLI outperformed GT. Killman (2014) tested GT, which remained an SMT system several years later, focusing on the quality of Spanish-to-English terminological output derived from this open-domain system in the legal context. A sample of over 600 terms and phrases was analysed from a collection of judgement summaries from the Supreme Court of Spain that exceeded 12,000 words. Nearly 65% of this sample was assessed as being rendered appropriately by GT. This result shows terminological reliability more often than not, despite challenges that may be associated with translating legal terminology and the fact that GT is a general-purpose system. Mileto (2019) reflected on an experience where she had students post-edit output from various SMT systems in a legal translation activity from Italian to English. The systems included the European Commission's MT@EC, replaced in 2017 by eTranslation (an NMT system), as well as SDL Language Cloud and GT within the SDL Studio workbench. Mileto (2019) argued that MT leads to productivity gains when combined with other tools or resources such as translation memories and glossaries.

DGT translators have been allowed to choose whether they prefer completing translation tasks with in-house MT. A couple of studies assessed their uptake of MT@EC (Cadwell et al., 2016; Rossi & Chevrot, 2019). Most of the translators in a focus group from all 24 language departments reported using MT every day and finding it useful (Cadwell et al., 2016). Translators who found it useful and not useful both asserted output quality as a primary motivator, while the former translators also pointed out speed or productivity increases. Translators from 15 language departments who were surveyed reported adoption rates that differed according to the language department, but which were overall high (Rossi & Chevrot, 2019). Their reasons were mostly to save time, but some of the translators surveyed reported seeking terminology suggestions or help with structural or grammatical aspects of text. Another study (Lesznyák, 2019) interviewed DGT translators from the Hungarian department and revealed that many of them found eTranslation useful for saving time, for inspirational purposes or to not have to translate from scratch. However, the majority of these translators expressed reservations. In the case of post-editing studies involving legal texts, the majority feature tailored MT. Farzindar and Lapalme (2009) conducted a pilot study involving post-edited output from TransLI (Gotti et al., 2008) in the translation of a judgement from French to English. Given how closely related this text was to the database of this statistical system, Farzindar and Lapalame (2009, p. 72) determined that "a cost-effective revisable text" can be produced. Macken et al. (2020) found that post-editing was faster than translating when they compared data obtained from translation and post-editing tasks completed by translators in the French and Finnish departments who post-edited output from MT@EC and eTranslation, respectively.

Post-editing was, on average, somewhat faster than translating in the case of MT@EC and eTranslation for both sets of translators, who rated the output similarly in both cases in the 3–4 range on a five-point scale. Post-editing in the case of eTranslation was also, on average, faster in a study with translators from the Polish department, though there was variance in speed among individuals (Stefaniak, 2020). Finally, a few studies focus exclusively on post-edited products produced by DGT translators using eTranslation. A couple of studies on output errors in German (Vardaro et al., 2019) and Slovene (Arnejšek & Unk, 2020) revealed errors in the following areas, among others: terminology, register, function words, polysemy and omissions. A study on post-edits completed in Dutch found that the DGT translators' edits were mainly related to register, style, or syntax (Desmet, 2021).

In the case of GT, several years after it transitioned to an SMT system, García (2010, 2011) and Şahin and Dungan (2014) conducted studies comparing post-editing output from this system with translating. In the case of García's studies, which involved English and Chinese legal and medical texts translated to and from these languages, the legal texts had the worst scores in two of the three sets of tests covered by these two studies (García, 2010, 2011). Nevertheless, participants' average scores were slightly better when these two texts were post-edited (García, 2011, p. 227). There was a minor increase in speed in one case and a minor decrease in the other (Ib., p. 223). In García's third set of tests (2011), the text with the highest average score happened to be legal, and post-editing resulted in a substantially higher average score (García, 2011, p. 227) and a minor increase in speed (García, 2011, p. 223). However, Şahin and Dungan (2014) found the quality to be slightly improved when the legal texts in their English–Turkish study were not post-edited by the participants but were instead translated when participants had access to the internet only. This result contrasts with that of technical texts in their study. No data on time was recorded with regard to these results.

The present study relates to and builds on the previous studies in different ways. Like the studies carried out by García (2010, 2011) and Şahin and Dungan (2014), it involved participants who prepared translations of passages of similar length by either post-editing GT output or translating. However, the tasks were carried out from English to Spanish, which was part of GT's first stage of supported languages that could be translated to and from English. The passages were legal in nature and provided by the American Translators Association (ATA). As in the case of García's studies (2010, 2011), the present study employed standardized scoring criteria from a translator accreditation authority, in this case the ATA. It relied on third-party evaluation of the translation products and it timed each translation task. The number of participants in the present study was a bit higher than the third and final set of participants in García's study (2011) and substantially higher than both his second set (García, 2011) and his first set (García, 2010, 2011), at nearly double in both cases. The number of participants in the present study was also more than double that in Şahin and Dungan's study (2014). In addition, both the post-editing and translating tests in the present study allowed participants to use internet resources, which aligns with a set of the tests conducted by Şahin and Dungan (2014) and García (2011). The details of the research design and methodology of the present study are provided below.

3 Methods

3.1 Participants

A volunteer convenience sample was recruited to participate in the study. Participation was limited to the geographical region of the institution where the study was conducted due to budget and logistics constraints. Recruitment was conducted through professional associations' directories, professional conferences and registries, and translator or professional networking websites.

The sample for this study included 26 translators who worked on a full- or part-time basis with 23% reporting full time translation work and the remaining 77% part-time. Experience varied from 1 to 30 years (*M*=9.8, *SD*=9.0), and all participants worked in the English–Spanish language combination: 50% translating mostly or exclusively from English to Spanish, 15% mostly from Spanish to English, and 35% working equally in both directions.

The participants also reported a range of language backgrounds and credentials. When asked to identify a native language in which the respondent was best at reading and writing, 50% reported Spanish and 11.5% reported

English. The remainder described themselves as having both Spanish and English as native languages, with 19% of the sample reporting that they were equally proficient in the two languages, 11.5% more proficient in English, and 8% more proficient in Spanish. More than half (69%) of the sample reported possessing a translation and/or interpreting credential, either academic or professional, while 31% reported no such credential.

There was a variety of reported experience with and attitudes toward MT. Among the participants, 57.7% reported prior use of MT in their work. Of those who had used MT, 81.25% reported choosing to do so voluntarily while the remaining 18.75% reported that a client, language service provider or translation agency had required use of a specific MT system. The 42.3% who did not report prior MT use expressed a variety of reasons for not using it, including not being required by clients, not having time to learn about MT, personal preference, and concerns about the accuracy and confidentiality of the process.

3.2 Procedures

Participants were recruited following approval by a university Institutional Review Board. The study took place in a language laboratory. Before taking part in the study, participants provided their informed consent and were informed that they were free to withdraw from the study at any time. In exchange for their time and voluntary participation, participants received a $50 gift card as compensation. All the enrolled participants completed the study.

The primary assigned tasks were to render two legal passages from English into Spanish: a workplace drug and alcohol policy and an order for release from custody. The passages, each of which are approximately 250 words in length, were provided by the ATA as one of the two domain-specific elective passages from which candidates choose. Participants were randomly assigned to translate one of these passages and post-edit the GT output of the other passage. Participants were allowed a total of three hours to complete both tasks plus a break between the two. While completing the tasks, they were allowed to use any internet resource they wanted, as long as it was not MT.

Additionally, participants responded to two questionnaires. Prior to the translation tasks, the questions were related to demographic variables and previous experience with MT. On completion of the translation tasks, additional survey questions were asked to give participants an opportunity to reflect on the utility of MT, and on aspects of the two passages.

3.3 Measures

The independent variable of principal interest is the experimental condition of translating versus post-editing. The analysis also controls for any potential difference in difficulty between the two texts, participant credentials, and experience. Participants were split into two groups based on whether or not they reported academic and/or professional translation and interpreting credentials (e.g., university translation and interpreting degrees and certificates or professional translation and interpreting certifications), and experience was divided between those who reported less and more than three years of experience. There are three dependent variables: time on task, quality score, and number of errors in the output.

During the translation task, Translog-II (Carl, 2012) was used to record completion time in addition to other process-oriented measures such as keystrokes that were not analysed in this study. The quality of the translations was assessed in two ways by an independent ATA-certified grader, for whom funding was available for up to 22 participants. The translations of the remaining four participants were assessed using the ATA grading rubric and by drawing on this grader's corpus of evaluations, which were plentiful enough at this stage to be used to consistently reproduce existing grading patterns. First, the grader employed the ATA standardized scoring criteria in which a lower score indicates fewer errors or improved quality. Standardized scoring criteria are used not only in professional certification contexts (Koby & Melby, 2013), but also as a means to standardize error marking (e.g., Koby, 2015)[3]. While the ATA framework has received some criticism

[3] ATA grading relies on a Flowchart for Error Point Decisions where errors are divided broadly between "mechanics/style" ("identifiable without viewing source text") and "transfer/strategy" ("identifiable by comparing to source text"), and points for errors in each of these categories may range from 0–4 and 0–16, respectively, depending on their severity (https://www.atanet.org/certification/how-the-exam-is-graded/error-points/). See Koby (2015) for a detailed review of the ATA flowchart.

in the literature, given the potential variability across multiple assessors (Phelan, 2017), the choice of using a standardized rubric by a single assessor was due to an effort to establish validity and reliability of the underlying construct (Angelelli, 2009), which in this instance is a quality metric. Second, the number of individual error occurrences in the translation were counted. The grader was not provided with any participant information and was blinded to whether the passages had been translated or post-edited. In the four cases mentioned above it was known how the participants translated each passage. However, there are sufficient grading examples in which knowledge about the experimental condition was determined not to have a demonstrable impact on the ability to consistently evaluate the quality of the translations using the ATA scoring rubric.[4]

3.4 Analysis

Data were analysed using ANOVA to assess the effect of the experimental condition while controlling for any effects due to differences between the two texts. The dependent variables were visually assessed for approximate normality in selecting parametric analysis. A post-hoc analysis with a Shapiro-Wilk test confirmed that all dependent variables approximately followed a normal distribution (Mellinger & Hanson, 2017). Possession of translation/interpreting credentials and experience (more or less than three years) were included as binary factors. An interaction effect was included for the condition and text variables. Because all independent variables were binary, no post hoc tests were necessary. Eta squared (η^2) was computed as an effect size with descriptions of the relative magnitude based on Cohen's (1988) figures, which are in line with Mellinger and Hanson (2017). Furthermore, post-task questionnaire data are included to analyse participant perceptions of MT. Some qualitative examples of GT's output and whether participants incorporated certain renditions into their post-edited version will be discussed.

4 Results

Prior to completing the translation tasks, participants were asked to respond on a 5-point Likert-type scale (from highly disagree to highly agree) to two items on whether MT has the ability to improve speed or quality. These questions assessed their expectations about performance when translating versus post-editing. For speed, the average response was 3.5 (*SD* = 1.3) with 62% either agreeing or strongly agreeing that use of MT can improve speed. Regarding quality, the average response was lower at 2.6 (*SD* = 1.1) with only 23% either agreeing or strongly agreeing that use of MT can improve quality. Overall, participants had a more favourable expectation of MT's ability to increase speed than to improve quality.

The first dependent variable of interest was total time on the task. An ANOVA with the four independent factors of condition, text, translation/interpreting credentials and experience showed that participants took longer to translate (*M*=67.1 minutes, *SD*=19.7) than to post-edit (*M*=56.6 minutes, *SD*=21.4). The interaction effect of condition and text was not statistically significant, which allowed for a focus on main effects. On average, participants took more than ten minutes longer to translate than to post-edit, though this effect was statistically significant only at the 10% level with a medium effect size ($F[1,46]=3.43$, $p=.070$, $\eta^2=.063$).[5] Holding a translation and/or interpreting credential was not a statistically significant factor with a small effect size; translators with such a credential (*M*=64.57 minutes, *SD*=20.36) took an average of more than 8 minutes longer to complete the tasks than those without such a credential (*M*=55.77 minutes, *SD*=22.03, $F[1,46]=1.96$, $p=.168$, $\eta^2=.036$). The results suggest that translators work faster when they are post-editing. Furthermore, the main effect for credentials implies that translators with translation and/or interpreting credentials spend more time on the task, regardless of whether they are post-editing or translating.

The second dependent variable is ATA score. The analogous ANOVA model to that estimated for time revealed that only the main effect for experimental condition was statistically significant with a large effect size. In particular, scores were higher for post-edited texts (*M*=21.9, *SD*=10.9) than for translated texts (*M*=34.7,

[4] For robustness, all analyses were conducted on the full dataset and after removing these four participants. The results were qualitatively similar, and the presented results utilize the full set of 26 participants.

[5] The three ANOVA models were estimated for the 26 participants measured under two conditions for a total of 52 observations. The models also included years of experience (more or less than three years) and possession of translation/interpreting credentials as additional variables. Finally, the models included an interaction term for text and condition. Therefore, the relevant *F*-statistics have five constraints, resulting in 46 degrees of freedom for the denominator.

SD=19.3, *F*[1,46]=8.31, *p*=.006, η^2=.148). Neither translation/interpreting credentials nor years of experience were significant factors in predicting quality for the participants. Therefore, the primary differentiator in this case is the experimental condition. This suggests that, on average, post-edited text is of higher quality.

The third and final dependent variable is number of errors since ATA flowchart error scores are variable according to type and severity of errors. The results of the ANOVA model echo the ATA score results with post-edited texts containing fewer errors. On average, post-edited texts contained 14.5 errors (*SD*=6.6) compared to translated texts with an average of 22.9 errors (*SD*=13.0, *F*[1,46]=8.92, *p*=.0054, η^2=.149). Again, neither translation/interpreting credentials nor years of experience were statistically significant in the model. These results reinforce the idea that post-edited output is, on average, of higher quality. The results of all three models are summarized in Table 1, which provides information on the tests for the difference between post-editing and human translation.

Table 1. Descriptive statistics and ANOVA results for the main effect of post-editing versus translation

Dependent variable	Post-editing	Translation	*F*[1, 46]	*p*
Time	56.6	67.1	3.43	.07
	(21.4)	(19.7)		
ATA score	21.9	34.7	8.31	.006
	(10.9)	(19.3)		
Number of errors	14.5	22.9	8.92	.005
	(6.6)	(13.0)		

Following the translation tasks, the participants responded to a number of additional questions. First, they responded to eight items on a 5-point Likert-type scale (see Table 2). The participants were generally quite favourable in their impression of the MT output, reporting beliefs that MT improved speed and quality. Agreement was also strong on the usefulness of the output (e.g., “contained good word and/or phrase translations” *M*=3.9). At the same time, participants felt strongly that the output required post-editing. For example, participants agreed that it contained grammar errors (*M*=4.0) and issues with word and phrase order (*M*=3.81).

Table 2. Post-task questionnaire results: counts, mean (M) and standard deviation (SD) on a 5-point Likert-type scale

Item	Highly disagree	Disagree	Neither agree nor disagree	Agree	Highly agree	Mean	SD
The MT output contained good word and/or phrase translations that I could use.	0	2	3	18	3	3.9	0.73
Translations of words and/or phrases were missing from the MT output.	0	3	8	14	1	3.5	0.76
The MT output contained grammar errors.	0	0	2	22	2	4.0	0.40
The MT output contained errors in the order of words and/or phrases.	0	1	6	16	3	3.81	0.69
The MT output helped me complete the finished translation product in less time.	2	4	1	12	7	3.70	1.26
The MT output contributed to the overall quality of the finished product.	1	5	4	14	2	3.42	1.03
The MT output slowed me down.	4	12	6	0	4	2.54	1.24
The MT output had a negative impact on the quality of my translation.	4	13	6	3	0	2.31	0.88

The post-task questionnaire also asked for perceptions of how the output aided or hindered translation speed. Participants were given several reasons why MT may have aided or hindered their performance and were asked to check all that applied. Table 3 provides the counts for each response. The reasons they were given relate in various ways to what DGT translators have reported as motivating their decision to use or not use MT or what aspects of MT output DGT translators report help or hinder them in areas such as speed, terminology or grammar (Cadwell et al., 2016; Lesznyák, 2019; Rossi & Chevrot, 2019; Mackenet al., 2020; Stefaniak, 2020), in addition to, for example, the terminological or omission errors found in MT output edited by DGT translators (Vardaro et al., 2019; Arnejšek & Unk, 2020) or the post-edits made in areas such as syntax (Desmet, 2020). Participants agreed that MT sped up their process by assisting with terminology, decreasing time spent thinking about how to correctly or adequately translate terms/phrases, and reducing time spent typing. Similarly, reasons why the participants believed MT hindered their performance included time spent considering what to use and time spent reading. Participants also noted that errors in MT output slowed down the process. The primary reasons selected, i.e., the order of words and grammar, relate to fluency being noted by Finnish DGT translators who worked with MT@EC as a primary issue (Macken et al., 2020) and to syntax being one of the main changes made by DGT Dutch translators to eTranslation output (Desmet, 2021).

Table 3. Post-task questionnaire results on how MT aided or hindered translation tasks

Question	Responses					
How did the MT output help you complete the finished translation in less time?	Cut down on the amount of time I had to spend looking up terms and/or phrases in dictionaries and websites.	Cut down on the amount of time I had to spend thinking about how to correctly translate a term and/or phrase.	Cut down on the amount of time I had to spend typing.	Provided alternatives to choose from.	Did not help.	
	13	13	18	1	3	
How did the MT output slow down the process?	Reading and understanding it slowed me down.	Considering what to reuse from it slowed me down.	I was slowed down by having to look up term and/or phrase suggestions in dictionaries and websites.	I was slowed down by having to fix the order of words and/or phrases.	I was slowed down by having to fix grammar errors.	I was slowed down by having to translate untranslated words and/or phrases.
	11	14	12	15	15	4

Finally, participants were asked to indicate for which textual features a proposed MT rendition was helpful. Again, multiple responses could be selected, which also relate to the aforementioned DGT studies in different ways. Help with terminology was the number one selection, followed by phrases, as shown in Table 4. These findings particularly relate to reasons stated by DGT translators for using MT@EC (Rossi & Chevrot, 2019). Three participants indicated that MT was not useful with any of the items. Overall, the results suggest that the perceived utility of MT in the translation of legal texts is primarily with the translation of particular words or phrases.

Table 4. Post-task questionnaire on useful aspects of MT

Item	Count
Terms	17
Phrases	10
Syntactic structures	7

Item	Count
Collocations	5
Grammar	3
Accent marks	1

Considered as a whole, the results show that, on average, post-editing was faster than translation and resulted in fewer errors and higher quality translations. The participants had generally favourable expectations and results from the post-editing task, and they listed different factors that might aid or hinder their speed and quality. Their perception of the utility of MT focused on terminology and phrases. In addition, in none of the analyses were years of experience statistically significant, which raises the question as to whether these results might generalize to a broad range of translators. Possession of translation/interpreting credentials was not statistically significant in lengthening time on the task. Regardless of the task, translators with such credentials took approximately eight minutes longer. This result ideally would be attributed to a commensurate increase in quality, but the data here cannot assess such a claim.

5 Discussion

The results of this study substantiate the pre-task majority belief held among the participants (62% either agreeing or strongly agreeing) that the use of MT could improve their speed, with the participants post-editing an average of ten minutes faster: an almost 16% decrease in time spent on task. An increase in speed is a particularly positive result when it does not come at the expense of quality, which in the case of this study was also observed as being enhanced by post-editing. This is so much so that the average quality score using the ATA scoring criteria (i.e., the lower the better) and number of detected errors decreased with statistical significance, even though only 23% of the participants prior to completing the tasks agreed or strongly agreed that using MT could improve quality. Post-editing reduced the average ATA score by as much as 47%, which is an arguably substantial reduction. These results are contrary to those collected by Şahin and Dungan (2014) in the legal translation component of their study. In the case of texts with the best and worst average scores in each of the three tests conducted by García (2010, 2011), which happened to be legal texts in each case, the results of the present study are reinforcing in terms of time and quality. Whereas García (2011, p. 223) reports a slight post-editing speed increase in two of these three legal texts, the speed increase in the present study is more than slight. In any event, the speed result in the present study is consistent with that in a couple of DGT post-editing studies (Macken et al., 2020; Stefaniak, 2020), even though the present study included output from a general-purpose SMT system. While García (2011, p. 227) reports moderate post-editing quality gains in the two tests where two of these legal texts had the worst average scores and a considerable quality gain in the test where the third such text had the best average score, the present study demonstrates considerable post-editing quality gains with statistical significance.

That the results in the present study differ from previous GT post-editing studies may be partly because the post-editing tasks involved a well-resourced language pair forming part of the first stage of development of GT. Moreover, the output was obtained when GT had been a statistical system for nearly a decade, so there had been ample time for improvements to be made, particularly since the system output was generated shortly before the underlying architecture transitioned to a neural system. In the case of the drug and alcohol policy text, for example, a couple of participants could really exemplify the potential of the output, in that they were able to complete the post-editing task in under 30 minutes (24.99 min and 28.03 min) with ATA passing scores (15 pts. in both cases). That 82.69% and 91.1% of the words in their finished products resemble the output displays how it could be quickly used without over-compromising quality.

Nevertheless, the two participants in these cases allowed erroneous aspects of output to creep into their final versions, amounting to more than half or all the error points they were marked. For example, the grader noted inconsistent translations of "company", which GT rendered as *compañía* in the first instance and *empresa* in other instances. Inconsistent translations provided by MT have been noted even in the case of institution-specific MT, such as eTranslation in the DGT context (e.g., Stefaniak, 2020). Likewise, they both engaged in unnecessarily changing the output in 23 instances in one case and nine in the other. This behaviour has been observed in previous studies examining translator behaviour when they revise translations (e.g., Mellinger

& Shreve, 2016) or post-editing MT output (e.g., Koponen & Salmi, 2017). The participant who did so in 23 instances incurred a few errors that could have been avoided by just accepting the corresponding MT suggestions. For example, this translator modified *un lugar seguro para sus empleados*, GT's rendition of "a safe place for its employees". Their changing it to *un lugar seguro laboral para sus empleados* led to the grader marking a syntax error, which could have been avoided with *un lugar laboral seguro para sus empleados*.

The profiles of these two participants are noteworthy, in that they had 13 and 15 years' experience, possessed no translation or interpreting credentials in one case, and credentials of both types in the other. Such profiles illustrate how translation/interpreting credentials might not be a reliable predictor of post-editing effectiveness. While in these two cases we might associate post-editing speed and quality with substantial years of experience (since their 13 and 15 years of experience are similar), we could refer to another participant with 25 years' experience, who also completed the same post-editing task with an ATA passing score (11 pts.) but in more than double the time (60.42 min). This participant's post-edited product also substantially resembles the output (82.82% of the words) with a comparable number of unnecessary edits (16 instances).

As the results show, participants were generally in agreement that post-editing contributed to improving speed and quality, which was supported by the general results in this study. Agreement was particularly strong that the output was useful at the terminological and phraseological levels. These are levels at which legal translation is often regarded as highly challenging (e.g., Šarčević, 1997; Alcaraz & Hughes, 2002, p. 47; Cao, 2007; Chromá, 2011; Biel, 2017) and where technological support might be of considerable value, especially if, as the participants widely agreed after completing the tasks, it reduces the amount of time spent looking up or thinking about how to translate terms and/or phrases.

Helpful MT suggestions included terms such as *cargo de delito mayor* (felony charge), *desacato* (contempt), *política* (policy), *programas de prestaciones* (benefit programs) and phrases or collocations such as *si usted no comparece* (if you fail to appear), *emitir una orden* (issue a warrant), *están obligados a cumplir con* (are required to abide by), or *ser castigado con pena de prisión* (be punished by imprisonment). These examples show how solutions from SMT can be contextually appropriate in cases where terminology or phraseology may entail lexical ambiguity or necessitate situationally variable target renditions (Killman, 2014). The first example (felony charge) is a system-bound term without a definite legal equivalent in Spanish, though the translation provided (*cargo de delito mayor*) is a common periphrastic if not coined rendition that is used. "Contempt" and "policy" are examples of terms whose Spanish translations may change depending on the context. For example, "contempt" in the general sense can be translated as *desprecio*, while *póliza* is an adequate translation of "policy" in the context of insurance. In the case of "benefits programs", "benefits" in the general sense is often translated as *beneficios*. In the case of the three phrases, translations of the verbs or verb phrases "fail to", "appear", "issue", "are required to" and "abide by" should be rendered in a way that is expected in the legal context. The omission of a translation of "fail to" is preferable to a forced rendition conjugating *omitir*, for example, while *comparecer* is the court-appropriate translation of "appear", unlike *aparecer*, which would be the common everyday translation. In the case of the verb "issue", *expedir* is a common translation but *emitir* collocates well with *orden* (warrant), and *están obligados a* collocates better than *se les requiere*, another possible rendition of "are required to". *Cumplir con* collocates better than *obedecer*, for example, since "abide by" precedes "the terms of this policy". Finally, *ser castigado con pena de prisión* is an idiomatic translation of "be punished by imprisonment", a less idiomatic rendition of which might be *ser castigado con encarcelación*.

While in most cases GT's output contained usable terms and phrases, there were several problematic renditions at this level, e.g., *convicta* (upon conviction), *se dará por terminado* (shall be terminated), *que se encuentre en violación de* (found in violation of), *no será elegible para la recontratación* (will not be eligible for rehire), and *usuarios* (users). More adequate translations might take the form of *una vez condenado* (upon conviction), *será despedido* (shall be terminated), *que infrinja* (found in violation of), *no podrá ser contratado de nuevo* (will not be eligible for rehire) and *consumidores* (users). This last case is an example of how the co-text "detect users and remove abusers of drugs and/or alcohol from the work force" is important to determine that "users" appears in the context of drugs and alcohol and not computers, for example. The remaining four cases exhibit how either prioritization of cognates or decomposition of phrases results in wording that is not (as) contextually desirable. In all but the first example (*convicta*), which more easily stands out as erroneous, these deficient renditions were accepted by post-editors in different cases.

6 Conclusion

The results of this study demonstrate substantial quality and speed gains when general-purpose MT is used with legal texts. They coincide with reported findings that SMT trained on corpora from a variety of domains produces a considerable level of quality in the legal domain (Koehn & Knowles, 2017). The participants in the present study were provided with usable translation suggestions, in numerous instances involving specialized terminology and phraseology, which at the same time often entailed lexical ambiguity or required translation solutions to be worded in specific ways. The participants' pretest majority belief that MT could help them improve speed (62% either agreeing or strongly agreeing) was matched by an increase in average time, that was not statistically significant. However, the pretest minority belief that MT could improve quality (23% agreeing or strongly agreeing) was countered by the results showing a significant increase in quality.

While the results of this study did not come from tests involving state-of-the-art NMT, it can be argued that they remain relevant given that results from NMT–SMT comparative studies are not entirely consistent. As previously pointed out in the case of the study by Koehn and Knowles (2017, p. 30), the SMT results, independent of whether they trained the NMT or SMT systems in their study on legal, medical, IT, Quran or subtitles corpora or on a mix of all these corpora, were evaluated in all cases as superior to varying extents when tested in the legal domain according to automatic metrics. Nevertheless, NMT and SMT were rated similarly by DGT translators who post-edited output from MT@EC and eTranslation in Macken et al. (2020). French was the language in the case of the statistical MT@EC and Finnish, with lower resources, was the language in the case of the neural eTranslation. Hence, the similarity of results may reflect the more recent progress made by NMT technology. Outside the legal translation context, NMT has been frequently determined as making gains in sentence fluency (Bentivogli et al., 2016; Bojar et al., 2016; Castilho et al., 2017b; Forcada, 2017, p. 305; Toral & Sánchez-Cartagena, 2017; Moorkens, 2018; Van Brussel et al., 2018; Stasimioti et al., 2020) but possibly to the detriment of accuracy (Castilho et al., 2017a), especially when the source text contains rare words or terminology (Sennrich et al., 2016; Wu et al., 2016). NMT accuracy improvements have also been found to different extents (Bentivogli et al., 2016; Castilho et al., 2017b; Van Brussel et al., 2018; Stasimioti et al., 2020)[6]. In any event, Koehn and Knowles (2017, p. 30) found that NMT and SMT systems trained exclusively on legal corpora performed better in the legal domain than their counterparts trained on other corpora. This finding echoes the study by Gotti et al. (2008), which demonstrates how an SMT system trained on a specific legal corpus can be highly useful when similar material is translated.

Post-editing tests could be carried out with NMT to determine if the results might vary from the results of the present study with respect to the impact of system use. Given that fluency improvements are more frequently cited in NMT, future post-editing studies might specifically test how post-editing legal texts might bear an influence in this regard. The peculiarity of morphology and syntax in legal texts has been cited in addition to terminology as a principal area of difficulty in legal translation (Alcaraz & Hughes, 2002, p. 18). Specific legal texts could be chosen to test productivity gains in these areas (e.g., statute laws, judicial rulings), and reveal individual morphological areas or syntactic challenges with which NMT might or might not help post-editors. In any event, it would be reasonable to posit at the current stage of MT development that participants might average higher levels of experience using MT and leverage the technology to a greater extent. Just over half (57.7%) of the participants in the present study reported prior use of MT in their work, with an overwhelming majority (81.25%) of them reporting having done so on a voluntary basis as opposed to being required in a professional assignment (18.75%). Furthermore, none of the participants with prior experience using MT reported post-editing on a regular basis. Given that MT has been making gains for some years following the SMT and NMT paradigm shifts, translators will likely have engaged more with using MT at this stage and may reveal more substantial speed and quality gains than in the present study.

[6] "Accuracy" is being used here to refer to a reduction in "lexical" errors (Bentivogli et al., 2016) and improvements in "adequacy" (Castilho et al., 2017b; Stasimioti et al., 2020; Van Brussel et al., 2018). Van Brussel et al. (2018, p. 3802) also find "lexical" errors, but they restrict these errors to the context of fluency. At the same time, they recognize possible overlap of lexical errors with semantic errors (Van Brussel et al., 2018).

Acknowledgments

This work was supported, in part, by funds provided by the University of North Carolina at Charlotte. The authors would like to thank Christopher D. Mellinger and Thomas A. Hanson for their substantial help with the statistical tests employed in this study and suggestions to tighten the analysis of the results. The authors would also like to thank the anonymous reviewers for their constructive feedback.

References

Alcaraz Varó, Enrique, & Hughes, Brian. (2002). *Legal translation explained*. Routledge.

Angelelli, Claudia V. (2009). Using a rubric to assess translation ability: Defining the construct. In Claudia V. Angelelli & Holly E. Jacobson (Eds.), *Testing and assessment in translation and interpreting studies: A call for dialogue between research and practice* (pp. 13–47). John Benjamins. https://doi.org/10.1075/ata.xiv.03ang

Arnejšek, Mateja, & Unk, Alenka. (2020). Multidimensional assessment of the eTranslation output for English-Slovene. In André Martins, Helena Moniz, Sara Fumega, Bruno Martins, Fernando Batista, Luisa Coheur, Carla Parra, Isabel Trancoso, Marco Turchi, Arianna Bisazza, Joss Moorkens, Ana Guerberof, Mary Nurminen, Lena Marg, & Mikel L. Forcada (Eds.), *Proceedings of the 22nd annual conference of the European Association for Machine Translation* (pp. 383–392). European Association for Machine Translation. https://aclanthology.org/2020.eamt-1

Bentivogli, Luisa, Bisazza, Arianna, Cettolo, Mauro, & Federico, Marcello. (2016). *Neural versus phrase-based machine translation quality: A case study* [Conference Presentation]. *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, *November 1-5*, *Austin, USA*. *CoRR*, abs/1608.04631. https://doi.org/10.18653/v1/D16-1025

Biel, Łucja. (2017). Researching legal translation: A multi-perspective and mixed-method framework for legal translation. *Revista de Llengua i Dret, Journal of Language and Law*, *68*, 76–188. https://doi.org/10.2436/rld.i68.2017.2967

Bojar, Ondřej, Chatterjee, Rajen, Federmann, Christian, Graham, Yvette, Haddow, Barry, Huck, Matthias, Jimeno Yepes, Antonio, Koehn, Philipp, Logacheva, Varvara, Monz, Christof, Negri, Matteo, Névéol, Aurélie, Neves, Mariana, Popel, Martin, Post, Matt, & Rubino, Raphael. (2016). Findings of the 2016 Conference on Machine Translation (WMT16). In *Proceedings of the First Conference on Machine Translation: Volume 2. Shared Task Papers* (pp. 131–198). Association for Computational Linguistics. https://aclanthology.org/W16-2301

Cadwell, Patrick, Castilho, Sheila, O'Brien, Sharon, & Michell, Linda. (2016). Human factors in machine translation and post-editing among institutional translators. *Translation Spaces*, *5*(2), 222–243. https://doi.org/10.1075/ts.5.2.04cad

Cao, Deborah. (2007). *Translating Law*. Multilingual Matters.

Carl, Michael. (2012). *Translog-II: A program for recording user activity data for empirical translation process research* [Poster session]. *13th International Conference on Intelligent Text Processing and Computational Linguistics*. 2012, New Delhi, India. https://research.cbs.dk/en/publications/translog-ii-a-program-for-recording-user-activity-data-for-empiri

Castilho, Sheila, Moorkens, Joss, Gaspari, Federico, Calixto, Iacer, Tinsley, John, & Way, Andy. (2017a). Is neural machine translation the new state of the art? *The Prague Bulletin of Mathematical Linguistics*, *108*(1), 109–120. https://doi.org/10.1515/pralin-2017-0013

Castilho, Sheila, Moorkens, Joss, Gaspari, Federico, Sennrich, Rico, Sosoni, Vilelmini, Georgakopoulou, Panayota, Lohar, Pintu, Way, Andy, Miceli Barone, Antonio V., & Gialama, Maria. (2017b). A comparative quality evaluation of PBSMT and NMT using professional translators. In *Proceedings* *of MT Summit XVI. Volume 1: Research Track, Nagoya, Japan* (pp. 116–131). https://aclanthology.org/2017.mtsummit-papers.10.pdf

Chromá, Marta. (2011). Synonymy and polysemy in legal terminology and their application to bilingual and bijural translation. *Research in Language*, *9*(1), 31–50. https://doi.org/10.2478/v10015-011-0004-2

Crego, Josep, Kim, Jungi, Klein, Guillaume, Rebollo, Anabel, Yang, Kathy, Senellart, Jean, Akhanov, Egor, Brunelle, Patrice, Coquard, Aurélien, Deng, Yongchao, Enoue, Satoshi, Geiss, Chiyo, Johanson, Joshua, Khalsa, Ardas, Khiari, Raoum, Ko, Byeongil, Kobus, Catherine, Lorieux, Jean, Martins, Leidiana, Nguyen, Dang-Chuan, Priori, Alexandra, Riccardi, Thomas, Segal, Natalia, Servan, Christophe, Tiquet, Cyril, Wang, Bo, Yang, Jin, Zhang, Dakun, Zhou, Jing, & Zoldan, Peter. (2016). Systran's pure neural machine translation systems. *CoRR*, abs/1610.05540. https://doi.org/10.48550/arXiv.1610.05540

Desmet, Luca. (2021). *An exploratory study of professional post-edits by English-Dutch DGT Translators*. [MA thesis]. Ghent University. [UGent Library repository].

Dik, Hugo. (2020). *CTRL+V for Verdict: An analysis of Dutch to English legal machine translation*. [MA thesis]. Leiden University. [Leiden University Student Repository].

Farzindar, Atefeh, & Lapalme, Guy. (2009). Machine translation of legal information and its evaluation. In Yong Gao & Nathalie Japkowicz (Eds.), *Advances in Artificial Intelligence* (pp. 64–73). Springer.

Forcada, Mikel L. (2017). Making sense of neural machine translation. *Translation Spaces*, *6*(2), 291–309. https://doi.org/10.1075/ts.6.2.06for

Forcada, Mikel L. (2020). Building machine translation systems for minor languages: challenges and effects. *Revista de Llengua i Dret, Journal of Language and Law*, *73*, 1–20. https://doi.org/10.2436/rld.i73.2020.3404

García, Ignacio. (2010). Is machine translation ready yet? *Target, 22*(1), 7–21. https://doi.org/10.1075/target.22.1.02gar

García, Ignacio. (2011). Translating by post-editing: Is it the way forward? *Machine Translation*, *25*, 217–237. https://doi.org/10.1007/s10590-011-9115-8

Gotti, Fabrizio, Farzindar, Atefeh, Lapalme, Guy, & Macklovitch, Elliott. (2008). Automatic translation of court judgments. In *Proceedings of The Eighth Conference of the Association for Machine Translation in the Americas*, 21–25 October 2008, Waikiki, Hawaii, USA (pp. 370–379). https://aclanthology.org/2008.amta-govandcom.11.pdf

Heiss, Christine, y Soffritti, Marcelo. (2018). DeepL traduttore e didattica della traduzione dall'italiano in tedesco. *InTRAlinea. Special Issue: Translation and Interpreting for Language Learners (TAIL)*. https://www.intralinea.org/specials/article/2294

Junczys-Dowmunt, Marcin, Dwojak, Tomasz, & Hoang, Hieu. (2016). Is Neural Machine Translation Ready for Deployment? A Case Study on 30 Translation Directions [Paper Presentation]. International Workshop on Spoken Language Translation, 8-9 December 2016, Seattle, WA, USA. https://doi.org/10.48550/arXiv.1610.01108

Kenny, Dorothy, & Doherty, Stephen. (2014). Statistical machine translation in the translation curriculum: Overcoming obstacles and empowering translators. *The Interpreter and Translator Trainer*, *8*(2), 276–294.

Killman, Jeffrey. (2014). Vocabulary accuracy of statistical machine translation in the legal context. In Sharon O'Brien, Michel Simard, & Lucia Specia (Eds.), *Proceedings of the Third Workshop on Post-Editing Technology and Practice (WPTP-3), The 11th Conference of the Association for Machine Translation in the Americas (AMTA)*, 22–26 October 2014, Vancouver, BC, Canada (pp. 85–98). AMTA. https://aclanthology.org/2014.amta-wptp.7.pdf

Kit, Chunyu, & Wong, Tak Ming. (2008). Comparative evaluation of online machine translation systems with legal texts. *Law Library Journal*, *100*(2), 299–321.

Koby, Geoffrey S., & Melby, Alan K. (2013). Certification and job task analysis (JTA): Establishing validity of translator certification exams. *Translation & Interpreting*, *5*(1), 174–210. https://doi.org/10.12807/ti.105201.2013.a10

Koby, Geoffrey S. (2015). The ATA Flowchart and Framework as a differentiated error-marking scale in translation teaching. In Ying Cui & Wei Zhao (Eds.), *The handbook of research on teaching methods in language translation and interpretation* (pp. 220–253). IGI Global. https://doi.org/10.4018/978-1-4666-6615-3.ch013

Koehn, Philipp. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X, 13–15 September 2005, Phuket, Thailand* (pp. 79–86). https://aclanthology.org/2005.mtsummit-papers.11.pdf

Koehn, Philipp. (2010). *Statistical machine translation*. Cambridge University Press.

Koehn, Philipp, Och, Franz J., & Marcu, Daniel. (2003). Statistical phrase-based translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, 27 May–1 June, Edmonton, Canada* (pp. 127–133). https://aclanthology.org/N03-1017.pdf

Koehn, Philipp, & Knowles, Rebecca. (2017). Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation of the Association for Computational Linguistics, 4 August 2017 28–39, Vancouver, BC, Canada* (pp. 28–39). https://aclanthology.org/W17-3204.pdf

Koponen, Maarit, & Salmi, Leena. (2017). Post-editing quality: Analysing the correctness and necessity of post-editor corrections. *Linguistica Antverpiensia*, *16*, 137–148. https://doi.org/10.52034/lanstts.v16i0.439

Lesznyák, Ágnes. (2019). Hungarian translators' perceptions of neural machine translation in the European Commission. In *Proceedings of MT Summit XVIII: Translator, Project and User Tracks, 19–23 August, Dublin* (pp. 16–22). https://aclanthology.org/volumes/W19-67

Lewis, Terence. (1997). When not to use MT and other translation tools. In John Hutchins (Ed.), *EAMT Workshop: Language Technology in your Organization?, 21–22 May 1997, Copenhagen, Denmark* (pp. 34–41). https://aclanthology.org/1997.eamt-1.5.pdf

Macken, Lieve, Prou, Daniel, & Tezcan, Arda. (2020). Quantifying the effect of machine translation in a high-quality human translation production process. *Informatics – Special Issue: Feature Paper in Informatics*, *7*(12), 1–19. https://doi.org/10.3390/informatics7020012

Mellinger, Christopher D., & Hanson, Thomas A. (2017). *Quantitative research methods in translation and interpreting studies*. Routledge. https://doi.org/10.4324/9781315647845

Mellinger, Christopher D., & Shreve, Gregory M. (2016). Match evaluation and over-editing in a translation memory environment. In Ricardo Muñoz Martín (Ed.), *Re-embedding Translation Process Research* (pp. 131–148). John Benjamins. https://doi.org/10.1075/btl.128.07mel

Mileto, Fiorenza. (2019). Post-editing and legal translation. *Digital Humanities Journal*, *1*(1). https://doi.org/10.21814/h2d.237

Moorkens, Joss. (2018). What to expect from neural machine translation: A practical in-class translation evaluation exercise. *The Interpreter and Translator Trainer*, *12*(4), 375–387. https://doi.org/10.1080/1750399X.2018.1501639

Phelan, Mary. (2017). Analytical assessment of legal translation: A case study using the American Translators Association Framework. *The Journal of Specialised Translation*, *27*, 189–210.

Roiss, Silvia. (2021). Y las máquinas rompieron a traducir… Consideraciones didácticas en relación con la traducción automática de referencias culturales en el ámbito jurídico. *TRANS. Revista de Traductología*, *25*, 491–505. https://doi.org/10.24310/TRANS.2021.v1i25.11978

Rossi, Caroline, & Chevrot, Jean-Pierre. (2019). Uses and perceptions of machine translation at the European Commission. *The Journal of Specialised Translation*, *31*, 177–200.

Şahin, Mehmet, & Dungan, Nilgün. (2014). Translation testing and evaluation: A study on methods and needs. *Translation & Interpreting*, *6*(2), 67–90.

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

Sennrich, Rico, Haddow, Barry, & Birch, Alexandra. (2016). Edinburgh neural machine translation systems for WMT 16. In *Proceedings of the First Conference on Machine Translation (WMT16)*, 11–12 August, Berlin, Germany (pp. 371–376). Association for Computational Linguistics. https://doi.org/10.18653/v1/W16-2323

Stasimioti, Maria, Sosoni, Vilelmini, Mouratidis, Despoina, & Kermanidis, Katia. (2020). Machine translation quality: A comparative evaluation of SMT, NMT and tailored-NMT Outputs. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation, Lisboa, Portugal* (pp. 441–450). European Association for Machine Translation. https://aclanthology.org/2020.eamt-1.47

Stefaniak, Karolina. (2020). Evaluating the usefulness of neural machine translation for the Polish translators in the European Commission. In André Martins, Helena Moniz, Sara Fumega, Bruno Martins, Fernando Batista, Luisa Coheur, Carla Parra, Isabel Trancoso, Marco Turchi, Arianna Bisazza, Joss Moorkens, Ana Guerberof, Mary Nurminen, Lena Marg, & Mikel L. Forcada (Eds.), *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation, 3–5 November, Lisbon* (pp. 263–269). European Association for Machine Translation. https://aclanthology.org/2020.eamt-1.28

Toral, Antonio, & Sánchez-Cartagena, Víctor M. (2017). A multifaceted evaluation of neural versus phrase-based machine translation for 9 language directions. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers* (pp. 1063–1073). Association for Computational Linguistics. https://aclanthology.org/E17-1100

Van Brussel, Laura, Tezcan, Arda, & Macken, Lieve. (2018). A fine-grained error analysis of NMT, SMT and RBMT output for English-to-Dutch. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC), Miyazaki, Japan* (pp. 3799–3804). European Language Resources Association (ELRA). https://aclanthology.org/L18-1600

Vardaro, Jennifer, Schaeffer, Moritz, & Hansen-Schirra, Silvia. (2019). Translation quality and error recognition in professional neural machine translation post-editing. *Informatics – Special Issue: Advances in Computer-Aided Translation Technology*, *6*(41), 1–29. https://doi.org/10.3390/informatics6030041

Wiesmann, Eva. (2019). Machine translation in the field of law: A study of the translation of Italian legal texts into German. *Comparative Legilinguistics*, *37*, 117–153. https://doi.org/10.14746/cl.2019.37.4

Wu, Yonghui, Schuster, Mike, Chen, Zhifeng, Le, Quoc V., Norouzi, Mohammad, Macherey, Wolfgang, Krikun, Maxim , Cao, Yuan, Gao, Qin, Macherey, Klaus, Klingner, Jeff, Shah, Apurva, Johnson, Melvin, Liu, Xiaobing, Kaiser, Łukasz, Gouws, Stephan, Kato, Yoshikiyo, Kudo, Taku, Kazawa, Hideto, Stevens, Keith, Kurian, George, Patil, Nishant, Wang, Wei, Young, Cliff, Smith, Jason, Riesa, Jason, Rudnick, Alex, Vinyals, Oriol, Corrado, Greg, Hughes, Macduff, & Dean, Jeffrey. (2016). Google's Neural Machine Translation System: Bridging the gap between human and machine translation. *CoRR*, https://doi.org/10.48550/arXiv.1609.08144

Yates, Sarah. (2006). Scaling the tower of babel fish: An analysis of the machine translation of legal information. *Law Library Journal*, *98*(3), 481–500.