informatics

*Article*

# Quantifying the Effect of Machine Translation in a High-Quality Human Translation Production Process

**Lieve Macken** [1,*], **Daniel Prou** [2] **and Arda Tezcan** [1]

[1] LT$^3$, Language and Translation Technology Team, Ghent University, 9000 Ghent, Belgium; Arda.Tezcan@ugent.be
[2] European Commission Directorate-General for Translation, 1140 Brussels, Belgium; Daniel.Prou@ec.europa.eu
* Correspondence: Lieve.Macken@ugent.be

Received: 27 March 2020; Accepted: 19 April 2020; Published: 23 April 2020

**Abstract:** This paper studies the impact of machine translation (MT) on the translation workflow at the Directorate-General for Translation (DGT), focusing on two language pairs and two MT paradigms: English-into-French with statistical MT and English-into-Finnish with neural MT. We collected data from 20 professional translators at DGT while they carried out real translation tasks in normal working conditions. The participants enabled/disabled MT for half of the segments in each document. They filled in a survey at the end of the logging period. We measured the productivity gains (or losses) resulting from the use of MT and examined the relationship between technical effort and temporal effort. The results show that while the usage of MT leads to productivity gains on average, this is not the case for all translators. Moreover, the two technical effort indicators used in this study show weak correlations with post-editing time. The translators' perception of their speed gains was more or less in line with the actual results. Reduction of typing effort is the most frequently mentioned reason why participants preferred working with MT, but also the psychological benefits of not having to start from scratch were often mentioned.

**Keywords:** machine translation; computer-aided translation; European Commission (DGT); post-editing; productivity

## 1. Introduction

The Directorate-General for Translation (DGT) of the European Commission translates written texts into and out of the EU's 24 official languages. As the largest institutional translation service in the world, DGT has a long tradition in using machine translation (MT). In the 1980s, a rule-based machine translation (RBMT) system (based on technology of Systran) was already operational [1,2]. A phrase-based statistical machine translation (PBSMT) system based on Moses [3], was introduced in 2013. From November 2017 onward, this system has gradually been replaced by a new system based on neural machine translation (NMT).

Machine translation is offered as a translation aid to all translators and is fully integrated in the translation workflow. At project creation (in a preprocessing step), the source text is sent to the MT engine and the MT output is stored as a translation memory exchange (TMX) file, the standard exchange format for translation memories. All DGT translators work within SDL Trados Studio in which basically two TMX files are imported: a TMX file which contains the retrieved matches from the huge central translation memory (EURAMIS) and a TMX file which contains the MT suggestions. Within SDL Trados Studio individual translators can then decide whether or not to use the MT suggestions for a given translation task.

In SDL Trados Studio, there are two different ways of interacting with MT: full segment mode or autosuggest. In the case of full segment mode, the MT suggestion is offered when there are no perfect or sufficiently high fuzzy matches found in the central translation memory. The autosuggest mode is the predictive typing function which suggests words or phrases in context based on what the translator is typing. SDL Trados Studio uses different colours to visualize perfect matches, fuzzy matches and MT suggestions, so translators are aware of the origin of the translation suggestions.

As is the case in many other organisations, Commission translators are under pressure to provide ever-larger volumes of translation with less and less resources while upholding very high quality standards. In this context, the vast majority of DGT's translators rely on machine translation, either occasionally or systematically, with the expectation that it will help them in their work. Internal MT use statistics show that, on average, translators choose to use machine translation in about 70% of cases when creating a new translation project. The overall aim of this study was to assess the actual impact of MT in DGT's translation workflow. More specific aims were to measure the speed gains (or losses) resulting from the use of MT by DGT translators in normal working conditions and to gain insight into the perceived usefulness of MT at DGT.

There are many scientific studies that focus on productivity in the context of post-editing machine translation output. Most of these studies, however, have been carried out in experimental conditions and not in real-life professional translation environments. To the best of our knowledge, this study is the first of its kind in the sense that it analyses data produced by highly professional translators while carrying out actual translation activities in their normal translation workflow. This study covers two language pairs (English–French and English–Finnish) and two different MT engines (a phrase-based statistical MT system for English–French and a neural MT system for English–Finnish). In total, data of 20 translators were collected over a period of one month at the end of 2018.

## 2. Related Research

Even though MT systems have made huge progress, MT quality is still variable, which means that in order to obtain high-quality publishable translations human translators still need to intervene. This intervention is generally referred to as post-editing. Moreover, in a context where translation quality requirements are extremely high (in the case of DGT for example translated texts are often legally binding [4]), even perfect MT suggestions have to be approved by human translators. Measuring the effect of MT thus basically boils down to measuring how humans interact with the MT suggestions, or, phrased differently, to measuring the human effort needed to adapt an MT suggestion to meet the required quality standards.

In his seminal work Krings [5] distinguishes three types of post-editing effort: temporal, technical and cognitive effort. Temporal effort is the time it takes to amend an MT suggestion to turn it into a high-quality translation. Technical effort refers to the amount of editing work that is involved in the post-editing process, and can be captured by the number of insertions, deletions and reorderings that are necessary. The third type of effort, cognitive effort, might be the most important one from a translator's point of view as it deals with the mental processes and cognitive load during post-editing.

As cognitive effort is difficult to measure studies often resort to a combination of temporal and technical effort. Technical effort can easily be derived from the post-edited segments by calculating the edit distance between the MT suggestions and the final post-edited segments. A metric that is often used is HTER [6], which takes into account insertions, deletions and substitutions of single words as well as shifts of word sequences. However, as various authors have argued, HTER only measures the shortest route to the final product but does not take into account all edits made during the process [7–9].

The three different types of post-editing effort are interrelated but not equal. Daems et al. [8] analysed the post-editing processes of student translators and professional translators, which were logged by using keystroke logging and eye-tracking, and found that different MT error types affect different post-editing effort indicators. Herbig et al. [10] argue that in addition to the technical and/or temporal effort, the actual cognitive load perceived by the post-editors should also be considered. Cognitive load may vary with individual post-editors and may be even to some extent be independent from MT quality. This is especially true for long sentences as was observed by Lesznyák [11]: "Translators see longer sentences as problematic not only because of the potential for lower NMT quality but because their complexity prevents a quick assessment of their correctness. It takes too much time and cognitive effort to analyse the components and decide what can be used."

Depending on the research context different methods can be used to measure the three different types of post-editing effort. In most experimental studies specially developed research environments are used. These environments not only register when a translation segment is opened and saved/closed, but as they are often connected to keystroke logging and eye-tracking devices, can provide very rich data sets of the post-editing process itself. Examples of such environments are PET [12], Translog-II [13] and Casmacat [14]. In such research environments, the precise time stamps of opening and saving/closing translation segments can be used to compute temporal effort. Keystroke logging tools produce more detailed information on translation speed and can also be used to measure technical effort by looking at all inserted and deleted characters and words. Moreover, they can also shed light on cognitive effort as pause analysis can be used to explore the underlying cognitive processes [15,16]. Eye-tracking tools measure eye positions and eye movements and are frequently being used in translation process research. Both fixation counts and fixation durations have been analysed as presumed indicators of cognitive effort [17–19]. Next to the more traditional indicators derived from keystroke and eye-tracking data, Herbig et al. [10] used a wider range of physiological sensor data including pupil dilatation, galvanic skin response, blood pressure, heart rate (variability) and respiration to estimate perceived cognitive load during post-editing.

It goes without saying that the research methods that are used in experimental studies, however valuable, cannot be used in real-world settings, as they would completely change the translators' normal way of working. Moreover, we fully agree with Läubli et al. [20] who argue that assessments of post-editing efficiency should be carried out in realistic translation environments. When the experimental conditions differ from the final working conditions in which post-editing will be used, the results most probably overestimate the obtained time gains through post-editing.

Although there are many experimental studies on post-editing MT, studies carried out in realistic environments are scarce and studies carried out with highly-qualified professional translators are even scarcer. Federico at al. [21] report on a field test carried out in SDL Trados Studio in which 12 professional translators were involved translating English texts from two domains (IT and legal) into German and Italian. They used the publicly available MyMemory-plugin which retrieves matches from the MyMemory translation memory server and MT suggestions from Google Translate in case no matches are found in the translation memory. All translators working on the same domain translated the same documents, half with only suggestions from the translation memory, half with suggestions from both the translation memory and Google Translate. Although the plugin also registered when a segment is opened and closed, not all obtained time measurements were reliable. As it is not possible to verify whether a translator is effectively working on an open segment, they defined two time thresholds and discarded segments with time measurements below the lower limit of 0.5 s per word and above the upper limit of 30 s per word. Furthermore, non-sequential translation processing posed problems for measuring the processing time for segments that were edited multiple times because the translator moved to a new segment without having completed and saved the current one. It total, they removed roughly 30% of translated words before carrying out analyses.

Läubli et al. [20] carried out a small-scale experiment in the translation workbench Across involving six translation students and four short texts (50 segments in total). As in the study of Federico et al. [21] they also compared two conditions: suggestions from the translation memory with and without MT suggestions. Screen recordings were used to obtain precise time measurements at text level.

Parra Escartín and Arcedillo [22] set up an experiment with 10 professional in-house translators translating an 8000-word software user guide from English into Spanish. MemoQ was used as translation environment as it keeps track of the time spent in each segment. A customized rule-based MT engine was used to generate the MT suggestions. As in the study of Federico et al. [21] they determined two time thresholds (an upper limit of 10 min per segment and a lower limit of 5 words per second) to remove outliers. They used the individual translation-from-scratch throughput as a reference to compare productivity gains for segments of different fuzzy match percentages with post-edited machine translated segments. They report extremely high translation-from-scratch throughput values and a large variation across individual translators, ranging from 473 to 1701 words per hour. The average productivity gain of post-editing machine translated segments was 24%.

Productivity is of course only part of the story. More qualitatively oriented research can shed light on the factors that affect the usefulness of MT. Several studies focus on the acceptance and perceived usefulness of MT at the European Commission. Cadwell et al. [23] used a focus group methodology in which 70 DGT translators participated and found an equally diverse set of reasons for using MT as for not using it. Language pair and the type of text to be translated were the two main reasons on the basis of which the translators decided to use MT or not. Terminological consistency was mentioned as a concern when using MT. Please note that in 2015, when the focus groups took place, all MT engines were still statistical.

Rossi et al. [24] interviewed 10 translators of the French DGT departments in 2017 and conducted a survey in 15 language departments afterwards with 89 responses. They found a high acceptance rate for MT usage and 37% of the translators mentioned speed gains as one of the reasons to use MT.

Lesznyák [11] carried out structured interviews with 38 Hungarian DGT translators and reported that the translators had widely divergent views on the usefulness of NMT. Incorrect and inconsistent terminology was also considered as a risk factor. Opinions on speed gains were mixed, but the reduction of typing effort was mentioned by half of the participants. She also highlighted the perceived psychological and cognitive benefits of not having to start a translation from scratch.

## 3. Method

### 3.1. Data Collection

In total, 20 professional translators participated in this study, 12 from the French and 8 from the Finnish language department at DGT. All participants were very experienced translators, each with at least 10 years of professional translation practice, and also experienced users of machine translation, each of them having used it professionally on a regular basis for one year or more. DGT translators are free to use MT or not and can decide whether or not to use it project by project. We therefore consider them 'regular' users and not 'systematic' users. Only a few DGT translators never use machine translation. However, the participants did not belong to this category. In the case of the participants translating into Finnish, all of them had used neural machine translation in their professional work for more than six months prior to the study; previously, they used statistical machine translation, the only technology available at the time. All 20 translators volunteered to participate in the study.

The 20 translators each took part in the experiment for a duration of one month. During this month, they worked in the same conditions as they usually do: they received actual work assignments they had to complete within the same deadlines and to the same quality standards that would have been the case outside the experiment. This setup ensured that the experiment would cover a representative sample of the document types and domains translated within DGT, while also ensuring that the quality standards of the final documents would be comparable to those produced outside experimental conditions. The bulk of documents translated within DGT are legislative documents and external communication documents, i.e., press releases and web site contents that cover a fairly wide variety of domains.

As already mentioned in the introduction, MT is fully integrated in the translation workflow at DGT. If the translator enables MT, in full segment mode, MT suggestions are available for all segments for which no perfect or sufficiently high fuzzy matches are retrieved from the translation memory. In autosuggest mode, the system suggests words or phrases in context based on what the translator is typing. If the translator decides to disable MT, translation suggestions are only retrieved from the translation memory and are thus only available for perfect or high fuzzy matches. All other segments are translated from scratch. During the data collection period, an English–French phrase-based statistical MT engine was used at the French language department (hereafter SMT-FR) as the neural MT engines for English–French were not yet used in production at that time; at the Finnish language department, all participating translators used the English–Finnish neural MT engine (hereafter NMT-FI).

In order to measure the impact of MT in this specific setting, only two types of segments are of interest: segments that were translated while an MT suggestion was available (either in full segment mode or autosuggest mode) and segments that were translated from scratch. To collect the data, we only minimally interfered in the normal translation workflow. We just asked the participants to enable MT for half of the segments in each document. For the other half of the segments in the same document, the translators were asked to disable MT and hence translate the source segments (for which no perfect or sufficiently high fuzzy matches are retrieved from the translation memory) from scratch. To control the impact of getting acquainted with the task at hand (with or without MT) and the broader textual context, MT was enabled or disabled for a different half of each consecutive document. In other words, if a translator enabled MT in the first half of a given document, they would enable MT in the second half of the next document and vice versa. After a document has been translated, the translators logged whether MT was used in the first or the second half of the document and the segment ID before which the MT had been enabled/disabled. This information was used to extract the segments of interest from the SDLXLIFF files and to analyse temporal effort.

The data we collected over a period of one month consisted of 186 XLIFF files, 101 documents from the French language department and 85 documents from the Finnish language department. In fact, more documents were translated during this one month period, but we decided to discard updated versions of source texts that had been translated earlier, as they would have provided no extra information. Table 1 presents the total number of segments, the total number of source words and the average number of source words per segment for the different segment types. No additional tokenisation has been performed prior to this analysis.

**Table 1.** Total number of segments (#s), total number of source words (#w) and average number of source words (#w (AVG)) per segment, for French and Finnish, per segment type.

| **Segment Type** | **French** | | | **Finnish** | | |
|---|---|---|---|---|---|---|
| | **#s** | **#w** | **#w (AVG)** | **#s** | **#w** | **#w (AVG)** |
| autoprop_edited | 25 | 359 | 14.36 | 430 | 722 | 1.68 |
| autoprop_unchanged | 1927 | 6513 | 3.38 | 1320 | 4620 | 3.50 |
| copy_source | 527 | 1264 | 2.40 | 1970 | 4866 | 2.47 |
| copy_source_edited | 354 | 6553 | 18.51 | 412 | 3869 | 9.39 |
| perfect | 1225 | N/A | N/A | 3322 | N/A | N/A |
| tm_edited | 1908 | 46,002 | 24.11 | 1449 | 30,067 | 20.75 |
| tm_unchanged | 2921 | 29,853 | 10.22 | 2555 | 21,258 | 8.32 |
| from_scratch | 1183 | 31,196 | 26.37 | 1165 | 25,921 | 22.25 |
| mt_edited | 1259 | 33,489 | 26.60 | 1271 | 31,406 | 24.71 |
| mt_unchanged | 107 | 984 | 9.20 | 75 | 792 | 10.57 |
| Total | 11,436 | 156,213 | 15.29 | 13,969 | 123,521 | 11.44 |

We can group the different types of segments, shown in Table 1 into three parts. In the upper section, we have the segment types automatically generated by the CAT tool by auto-propagation or by copying the source segment to target. The middle section consists of the segment types in which the translations are retrieved from the translation memory. The bottom section consists of the three segment types of interest in this study: segments that were translated while an MT suggestion was available (referred to as mt_edited when the MT output had been modified and as mt_unchanged otherwise) and the segments that were translated from scratch (from_scratch). In the analyses, we only focused on these three types of segments and ignored all other segment types. Please note that the segments of interest represent 22% of all segments and 42% of all translated source words for French and 18% of all segments and 47% of all translated source words for Finnish. The segment type 'perfect' is not included in the total number of words as such segments were not directly visible in the SDLXLIFF files.

By definition, the temporal effort involved in both tasks can be measured by the time it takes to achieve a high-quality translation, either by translating a source segment from scratch or by adapting an MT suggestion (for temporal effort the full segment mode and autosuggest mode segments were analysed together). For each segment, we measured the processing speed for both tasks in seconds per source word. While this can be considered a straightforward measurement, extracting reliable time measurements from the SDLXLIFF files proved to be more challenging than originally expected.

A first challenge is related to the computer-assisted translation tool that was used. The SDLXLIFF files only contain time stamps of the closing of segments after the last modification (the modified_on time stamp) but do not log when segments are opened for editing (instead, SDL Trados Studio stores a time stamp for when a segment is created in a translation memory for the first time). So, the only way to obtain reliable time measurements from the SDLXLFF files was to compare the segment's modified_on time stamp with the same time stamp of the previous segment. However, this way of working can of course not be used when translators process the text non-sequentially, which happened relatively frequently in our data set, as can be seen in Table 2. Translators not only edited segments non-sequentially and multiple times over different time periods, they also adopted very creative translation procedures. Some of them even used regular expressions to speed up their work. For that reason, we could only extract reliable processing times for segments that were translated sequentially by comparing the segment's modified_on time stamp with the modified_on time stamp of the previous segment provided that no other segment was edited between the two time stamps.

A second challenge was the lack of control we had over potential interruptions that could take place during the translation work (such as taking breaks, answering phone calls, switching to more urgent tasks) during which a segment could be left open without actually working on it. In a similar study that analyzes temporal effort involved in translation and post-editing tasks, Federico et al. [21] introduced two threshold values. They argue that processing times above 30 s per word are assumed to be due to software errors or the translator's behaviour (pauses, distractions, etc.), and processing times below 0.5 s per word due to accidental interactions with the software (e.g., saving a segment without reading or editing it). In order to exclude measurements that are most probably not related to the complexity of the task at hand, we filtered out segments with processing times outside these two threshold values.

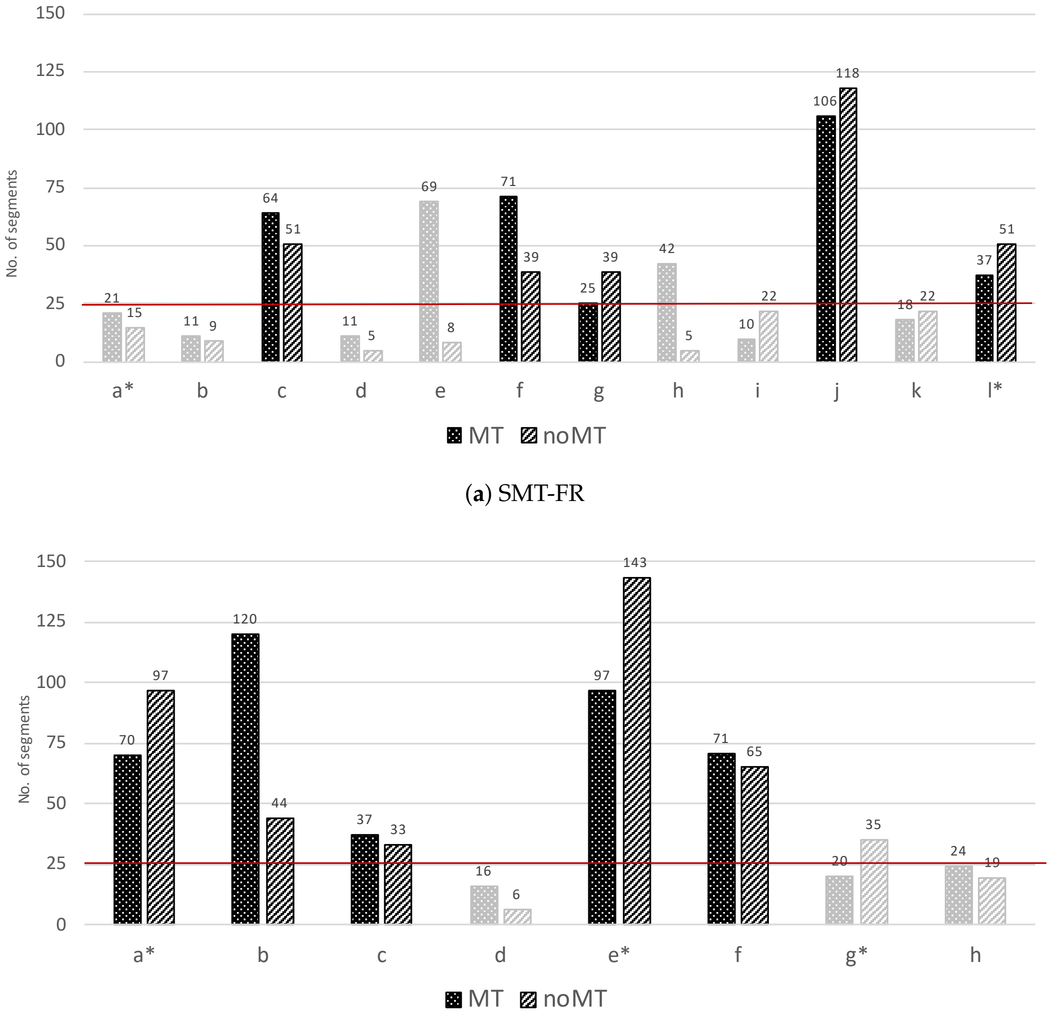
Finally, in the SDLXLIFF files, we also encountered a number of segments that did not contain the expected valid metadata values that were necessary to measure processing speed, such as the segment ID or the modified_on time stamp. Therefore, we also discarded these segments in our analyses. The number of discarded segments is given in Table 2; the number of retained segments per translator is visualised in Figure 1.

Using the above-mentioned filtering techniques (which removed non-valid segments, non-sequentially translated segments and segments with processing times below and above the time thresholds, in the given order, we retained a total of 869 segments and 18,343 source words for the French and 897 segments and 16,027 source words for the Finnish data set for further analyses.

Please note that applying the filtering methods in different order can result in different number of segments filtered out by each filter. Table 2 shows the number of segments and source words that were retained and the number of segments and source words that were removed by the different filters. Figure 1 further shows the total number of segments retained per segment type, per translator.

**Table 2.** The number of segments (#s), words (#w) and average number of words (#w (AVG)) filtered out by different methods for French and Finnish, after applying the filters in the given order.

| **Filter Type** | **French** | | | **Finnish** | | |
|---|---|---|---|---|---|---|
| | **#s** | **#w** | **#w (AVG)** | **#s** | **#w** | **#w (AVG)** |
| Total | 2549 | 58,721 | 23.03 | 2511 | 50,299 | 20.03 |
| Non-valid segment ID | 4 | 108 | 27 | 9 | 242 | 26.88 |
| Non-sequential editing | 1400 | 34,300 | 24.5 | 1202 | 25,884 | 21.53 |
| Non-valid time stamp | 133 | 2787 | 20.95 | 188 | 3153 | 16.77 |
| Processing speed threshold | 147 | 3291 | 22.38 | 223 | 5235 | 23.48 |
| Retained no. of segments | 869 | 18,343 | 21.10 | 897 | 16,027 | 17.87 |



(**a**) SMT-FR

(**b**) NMT-FI

**Figure 1.** Total number of segments retained per translator at (**a**) The French and (**b**) the Finnish language departments, per segment type. The translators who worked in auto-suggest mode are marked by a '*'.

As can be seen from Figure 1, the total number of retained segments per segment type varies among translators, from a minimum of 16 (translator d in the French department) to a maximum of 240 (translator e in the Finnish department). Moreover, the number of available segments per segment type is rather unbalanced. In the analyses, we used all available segments (869 for French and 897 for Finnish) to calculate average processing speed (i.e., we average over all translators). However, in order to calculate processing speed for individual translators, we only retained the translators for which we had a minimum of 25 segments per task (translators c, f, g, j and l of the French department, and a, b, c, e, f of the Finnish department). This threshold is highlighted in the corresponding charts with a red line. Next to temporal effort, we also calculated technical effort for all MT segments as we wanted to examine whether a correlation could be found between the two. We used human-targeted translation edit rate (HTER) [6], which measures the minimum number of edits that are required to transform the MT output into the final translation. Edits are defined as insertions, deletions, substitutions or reordering of word sequences and the HTER score is calculated as the total number of edits divided by the number of words in the final translation. As such, a low HTER score indicates a few number of edits and minimal technical effort involved. We additionally used CharacTER [25], which similarly to HTER, calculates the minimum number of edits required to adjust the MT output so that it matches the post-edited translation, normalized by the length of the MT output. However, unlike HTER, which works at word level, CharacTER measures edits at character level. CharacTER, therefore, allows us to make a distinction between 'heavy' edits (such as substituting one word with another) and 'light' edits (such as modifying only a suffix), and might also be better suited for morphologically rich languages such as Finnish. Given that HTER and CharacTER require two versions of a translation to measure the number of edits, they can not be used to measure the effort that was needed to translate from scratch.

### *3.2. Survey*

At the end of the one-month logging period the experiment was concluded by means of a survey in which we asked the participants how they perceived their own translation speed when working with and without MT, whether they preferred working with or without MT, and what their general impression of MT quality was. We asked the following questions:

- When making use of machine translation suggestions, do you think you work slower/at the same speed/somewhat faster/much faster than without using machine translation?
- Has participating in the study, during which you were partially prevented from using machine translation, changed this perception?
- Time constraints aside, do you prefer to translate with or without machine translation? Why?
- Is there a difference in that regard depending on document type?
- What is your assessment of the quality of the machine translation output you are provided with in DGT, as related to your professional translation needs: very poor, poor, ok, good, excellent?
- From a professional perspective, what are the main problems in the machine translation suggestions you are provided with?

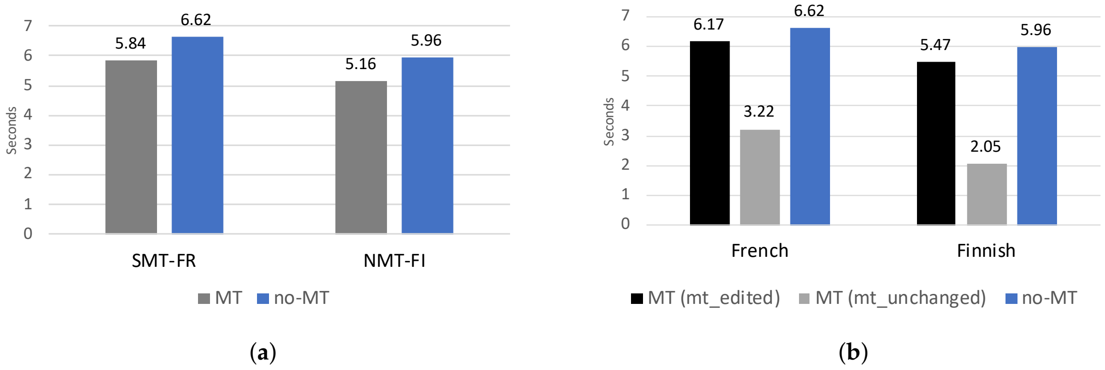
## 4. Results

In this section we first report the results on processing speed (Section 4.1) and then examine the relationship between processing speed and amount of editing (Section 4.2). We end this section by presenting the results of the survey (Section 4.3).

### *4.1. Processing Speed*

The primary goal of this study was to measure the speed gains (or losses) resulting from the use of MT by DGT translators in normal working conditions and to assess the potential benefits of enhancing a CAT tool with MT. We first calculated average processing speed per source token for all segments that were translated from scratch (no-MT) and all segments that were translated when an

MT suggestion was available (MT). The number of source tokens were calculated after tokenising the source segments. Tokenisation involves separating word from punctuation marks. In Figure 2 we provide this information per target language. Please recall that in the French language department a phrase-based statistical engine was used (SMT-FR), whereas in the Finnish department the MT engine was neural (NMT-FI). In Figure 2b we make a further distinction between MT suggestions that were modified (mt_edited) and MT suggestions that were accepted without editing (mt_unchanged).



(**a**) (**b**)

**Figure 2.** Average processing speed (expressed in seconds per source token) for all translators, per segment type and per language department.

As shown in Figure 2a, the average processing speed for the MT segments was lower than for the no-MT segments, both for SMT-FR and NMT-FI. The average speed gain when using MT was 12% for SMT-FR and 14% for NMT-FI. Figure 2b further shows that even when a given MT suggestion can be accepted without any modifications (mt_unchanged), the translators still spend on average 3.22 and 2.05 s per source token for SMT-FR and NMT-FI respectively. This is not surprising as the translators need time to read and approve even perfect MT suggestions. These values can be considered as upper boundaries for potential speed gains that can be expected at DGT under the assumption that MT produces perfect translations. When provided with such perfect MT suggestions, the translators at the French and Finnish departments work on average 51% and 66% faster, compared with translating from scratch.

While Figure 2 gives us a broad idea about the usefulness of MT at DGT, these results do not tell us whether these speed gains are obtained by all individual translators. In order to examine individual differences, in Figures 3–6, we show the average processing speed and the relative speed gains (or losses) when using MT, for the translators from which we retained at least 25 segments per task. Figure 3 shows the average processing speed for the translators of the French language department.

There are a two important observations that we can make based on the results presented in Figure 3. Firstly, we see a large variation in average processing speed among individual translators. The average translation speed (from scratch) varies between 5.01 s (translator f) and 8.02 s (translator g) per token (no-MT). With MT, the fastest translator spends on average 3.96 s (translator f), whereas the slowest translator spends 8.76 s (translator l) per source token. Such large individual differences are not entirely unexpected and have been reported in earlier research [22]. Secondly, Figure 3 shows us that, even though the availability of MT seems to lead to speed gains for the majority of the translators (three out of five), it leads to speed losses for some, namely for the translators g and l. In Figure 4, we provide the relative speed gains (or losses) per translator.

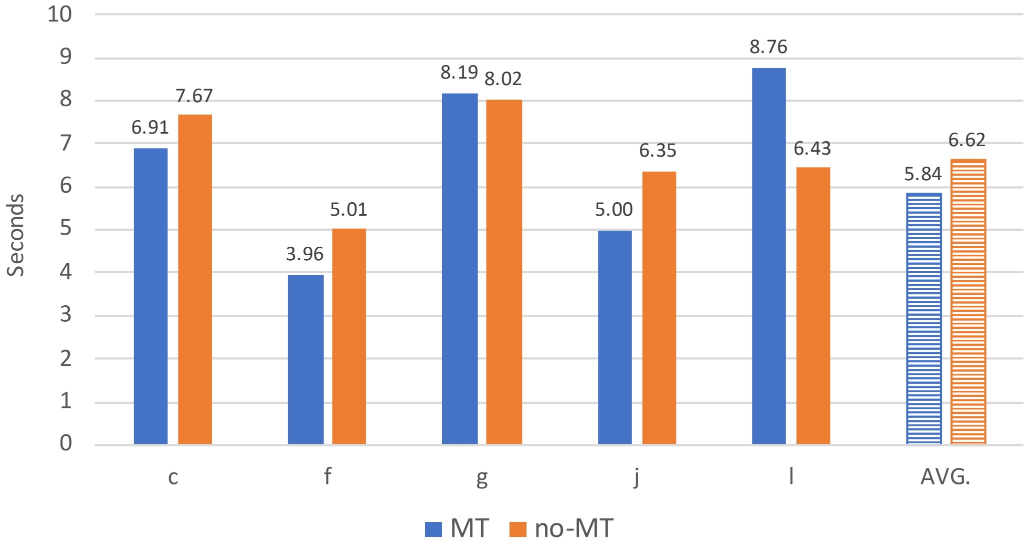


**Figure 3.** Average processing speed per task (expressed in seconds per source token) for five translators from the French language department. 'AVG.' stands for the average processing speed per task for all translators of the French language department.

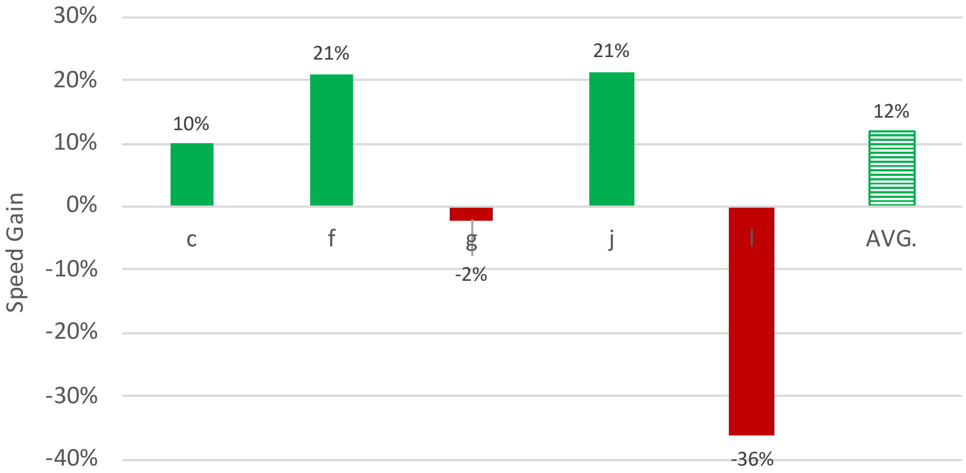


**Figure 4.** Relative speed gains when using MT, for five translators from the French language department. 'AVG.' stands for the average speed gain for all translators of the French language department.

The speed gains (displayed as positive values) for the three translators out of five range from 10% (translator c) to 21% (translators f and j). For the two remaining translators enabling MT seems to slow down their work, as we observe speed losses (displayed as negative values) ranging from 2% (translator g) to 36% (translator l) .

In Figures 5 and 6, we show the results for the Finnish translators. Figure 5 shows us that enabling MT leads to speed gains for all five Finnish translators. These results also show less variation in processing speed among translators, compared to the French department. When translating segments from scratch (no-MT), the processing speed per source token varies between 5.34 s (translator e) and 6.81 s (translator b). When MT is enabled the fastest translator (translator c) spends on average 4.58 s per source token and the slowest translator (translator b) 5.53 s.

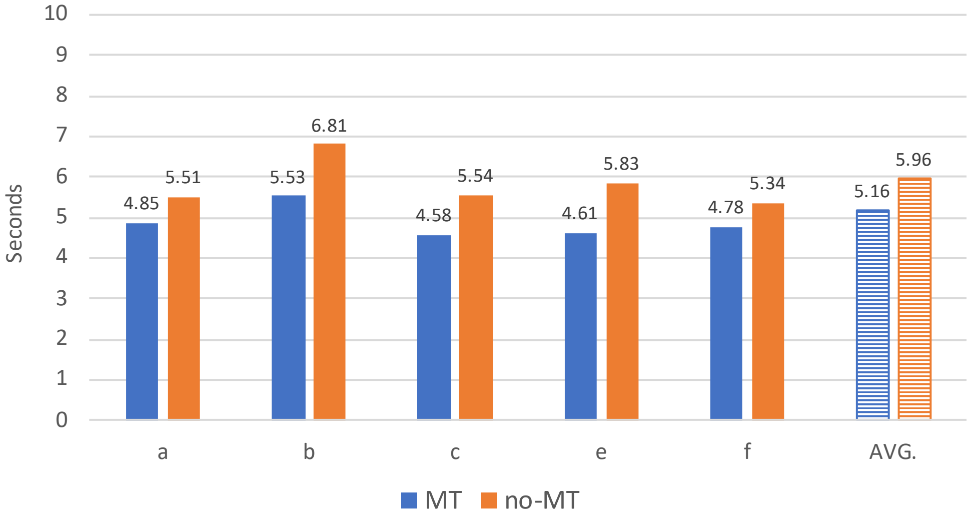


**Figure 5.** Average processing speed per task (expressed in seconds per source token) for five translators from the Finnish language department. 'AVG.' stands for the average processing speed per task for all translators of the Finnish language department.

In Figure 6, we see the relative speed gains (or losses) for five translators from the Finnish language department.

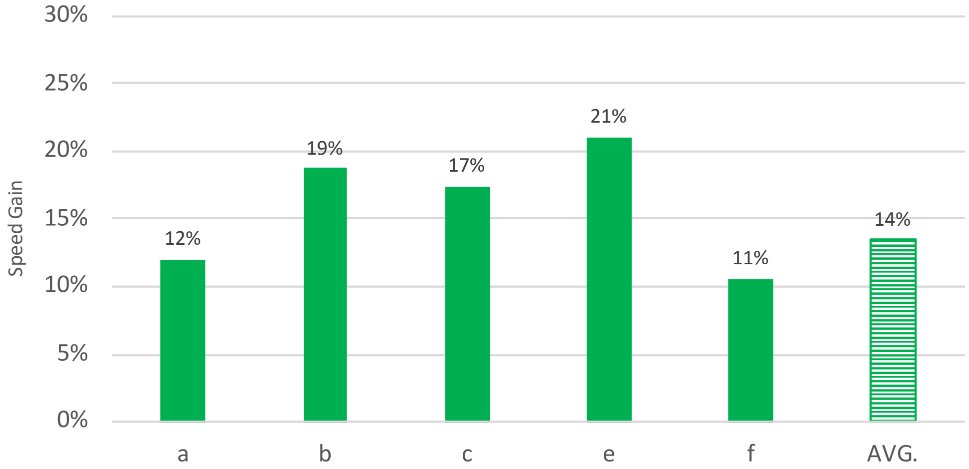


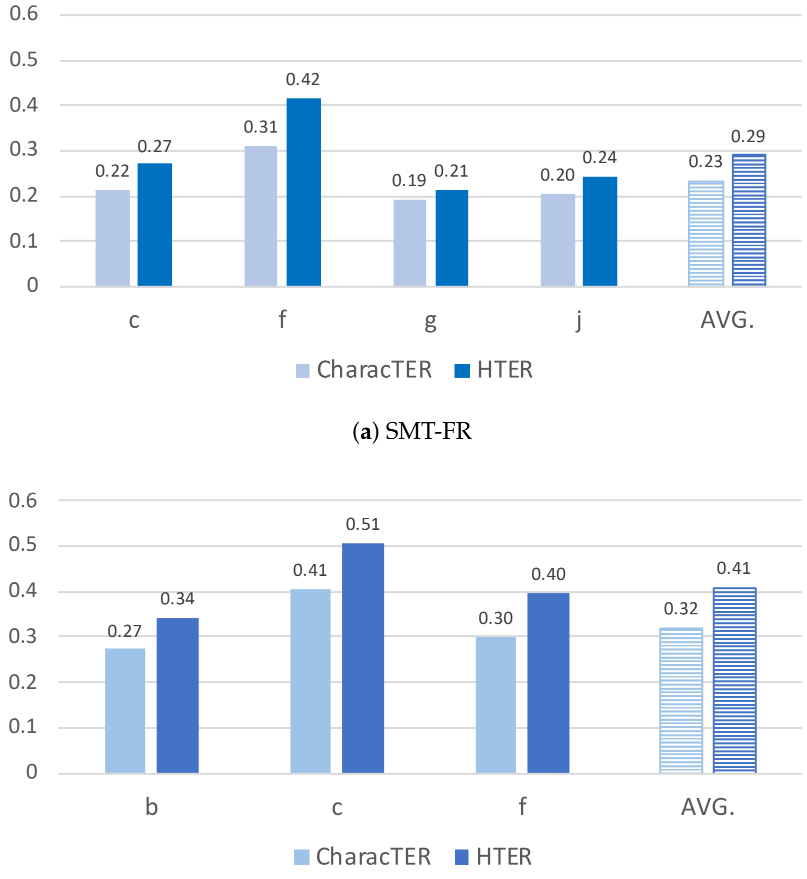
**Figure 6.** Relative speed gains when using MT for five translators from the Finnish language department. 'AVG.' stands for the average speed gain for all translators of the Finnish language department.

All translators, in this case, achieve a speed gain of minimum 11%. These results also point to a smaller difference between the minimum and maximum relative speed gains compared to the French translators. For Finnish, we observe relative speed gains ranging from 11% to 21%.

### *4.2. Amount of Editing*

A secondary goal of this study was to analyse how much editing was actually done on the MT suggestions (i.e., technical effort) and how the amount of editing relates to post-editing speed (i.e., temporal effort). More precisely, we were interested in whether there is a correlation between technical and temporal effort measurements, which, if there is, would enable us to use technical effort measurements as an approximation for temporal post-editing effort. This analysis has only been carried out for the segments that were post-edited using the MT suggestions in full segment mode as it is impossible to reconstruct the MT suggestions that were available in autosuggest mode.

In Figure 7, we report both the word-level HTER and character-level CharacTER scores (on segment level) for a subset of translators from the two language departments we analysed in the previous section and who post-edited MT output using the full segment mode. In the same figure, we also report the average scores, which were calculated by using all the segments that were post-edited using the full segment mode, for each language department.



(**a**) SMT-FR

(**b**) NMT-FI

**Figure 7.** CharacTER and HTER scores per translator at (**a**) the French and (**b**) the Finnish language departments.

As higher CharacTER and HTER scores correspond with higher number of edits, Figure 7 shows us that, on average, more editing occurred for NMT-FI than for SMT-FR, with average CharacTER scores of 0.32 vs. 0.23 and HTER scores of 0.41 vs. 0.29, for the Finnish and French translators, respectively.

In order to have clear understanding about the relationship between technical and temporal post-editing effort, we measured linear (Pearson's correlation coefficient) and any type of monotonic (Spearman's correlation coefficient) relationship between post-editing time (per word) and CharacTER and HTER for the French and Finnish translators. The results are provided in Table 3.

**Table 3.** Correlation between temporal post-editing effort (post-editing time, PET) and technical effort measures (HTER and CharacTER) per language department.

| | **PET/HTER** | **PET/CharacTER** |
|---|---|---|
| **SMT-FR** | | |
| Pearson ($r$) | 0.157 | 0.151 |
| Spearman ($\rho$) | 0.327 | 0.328 |
| **NMT-FI** | | |
| Pearson ($r$) | 0.305 | 0.306 |
| Spearman ($\rho$) | 0.401 | 0.396 |

While extreme coefficient values of less than 0.1 (indicating a negligible correlation) and greater than 0.9 (indicating a very strong correlation) are easy to interpret, interpretation of the values in-between are often disputable [26]. Nevertheless, based on the conventional interpretations used in literature (see also [26] for a brief overview and discussion), we can conclude that, for SMT-FR, post-editing time (PET) does not have a strong correlation with HTER or CharacTER measurements, neither with respect to $r$ or $\rho$ ($r_{HTER} = 0.157$, $p = 0.005$; $r_{CharacTER} = 0.151$; $p = 0.006$, $\rho_{HTER} = 0.327$, $p < 0.0001$; $\rho_{CharacTER} = 0.328$, $p < 0.0001$). On the other hand, the $\rho$ values are higher than $r$ values, which indicates that the relationship between temporal and technical post-editing effort is non-linear. Not surprisingly, for the French translators, the CharacTER and HTER values correlate strongly with each other, with respect to both correlation measures ($r = 0.877; \rho = 0.911$, with $p < 0.0001$ for both values).

The results for the Finnish translators show roughly similar patterns. Post-editing time does not correlate strongly with HTER and CharacTER measurements ($r_{HTER} = 0.305$; $r_{CharacTER} = 0.306$; $\rho_{HTER} = 0.401$; $\rho_{CharacTER} = 0.396$, with $p < 0.0001$ for all values). On the other hand, both types of correlations seems to be somewhat stronger compared to the results for the French translators. Finally, it seems that using CharacTER (a character-based measure) for measuring post-editing effort for the Finnish language, which is morphologically richer than French, does not necessarily provide us with a different perspective on technical post-editing effort than using HTER (a word-based measure). Similar to the results for the French department, for the Finnish department, the $r$ and $\rho$ values correlate strongly with each other ($r = 0.836, \rho = 0.889$, with $p < 0.0001$ for both values).

### *4.3. Survey*

All translators filled in the survey. The first set of questions were intended to get insight into perceived speed gains or losses when using MT. Most translators think they work faster or much faster when using MT (see Table 4). Only three translators (translators j and l of the French language department and translator a of the Finnish language department) think they work at the same speed. When we look at the actual speed gains or losses of these three translators in Figures 4 and 6, we see some discrepancies. The French translator l works not at the same speed but actually 36% slower; the French translator j works 21% faster and the Finnish translator a works 12% faster.

**Table 4.** Answers to the first question: When making use of MT suggestions, do you think you work slower/at the same speed/somewhat faster/much faster than without using MT?

| | **Slower** | **Same Speed** | **Somewhat Faster** | **Much Faster** |
|---|---|---|---|---|
| SMT-FR | 0 | 2 | 7 | 2 |
| NMT-FI | 0 | 1 | 7 | 1 |

The comments of the participants help to put the results into perspective.

- One participant notes that the use of machine translation only has a very marginal effect on the overall speed: "Other factors have a much greater impact, such as the clarity of the text, its technical complexity, its links with previous texts, its degree of idiomaticity, and how much research is necessary to understand it and choose the right terminology. In other words, MT might save me some of the time necessary to type the text, but typing is a very small fraction of the work involved in the translation of a text."
- One participant relates speed to both translation memory match rate and the quality of the original documents: "With a low (retrieval) match rate, I work somewhat faster with MT than without if the original is in English and well written. With higher match rates, I work at the same speed. With badly written originals, I might even work much slower (as I use full segment insertion, I have to delete text then retype my translation)."

In the second set of questions we asked whether the participants preferred working with or without MT and to motivate their answers. The reduction of typing effort was explicitly mentioned by seven participants and was as such the most frequently given reason why the participants preferred working with MT. The second most frequent reasons are what Lesznyák [11] calls the psychological benefits. In the survey we read the following candid comments:

- "I prefer to translate with machine translation because I have the impression that I have a *base*, a foundation on which I can build. It's a reassuring feeling."
- "Not having to start from scratch is reassuring."
- "I also like the feeling that I'm not working alone—even if it is just a silly machine that is there to *help* me."
- "With MT, because I am lazy, and when I see that the segments are already filled, it gives me the impression that the work is already partly done."

Other reasons why the participants preferred working with MT were that it saves time as a translator does not have to look up that many words anymore in a dictionary ("especially for words that you know, but don't remember"); the usage of MT also decreases the probability of leaving out some elements; it sometimes provides you with solutions you would not think of at first. One participant mentioned that they preferred working with MT for long, repetitive and/or boring texts and preferred working without MT for more creative/recreational translations. Reasons for working without MT were that MT sometimes orientates you in wrong directions and that it made translators less critical. With respect to document types, MT is perceived less useful for more technical texts and technical annexes, normative documents (e.g., model contracts, model letters, templates, legislation building on a previous body of legal texts like amending acts, treaties) where there is plenty of previous normative material to resort to, and documents with a lot of elements that should not be translated such as numbers, proper names or website addresses, for which MT tends to make "slightly crazy" suggestions.

The last question dealt with MT quality. As can be seen in Table 5, the results for SMT-FR and NMT-FI were very similar. Only one participant in each language department assessed MT quality as poor; all other participants gave either the score OK or Good.

**Table 5.** Answers to the last question: What is your assessment of the quality of the machine translation output you are currently provided with in DGT, as related to your professional translation needs: very poor, poor, OK, good, excellent?

| | Very Poor | Poor | OK | Good | Excellent |
|---|---|---|---|---|---|
| SMT-FR | 0 | 1 | 4 | 6 | 0 |
| NMT-FI | 0 | 1 | 3 | 5 | 0 |

Poor grammar, word order, and other grammatical problems such as plurals, gender and conjugations were mentioned as main problems in SMT-FR. This is not entirely unexpected as fluency problems are typical for phrase-based statistical systems. Other issues that were mentioned were words that are not adequate in the context, repetition of a wrong translation throughout the document, poorly interpreted originals.

For NMT-FI most issues are related to accuracy. The NMT system sometimes produces complete nonsense. A new (minor) problem is that the NMT system sometimes creates compound words which do not exist as such in Finnish. However more problematic are the less obvious mistakes, as one participants phrased it: "It is fluent but might at the same time be hiding serious mistakes that you only spot when you start looking deeper." Apart from accuracy problems, the MT also does not recognise the style and register needed.

## 5. Discussion

The overall aim of this study was to assess the impact of MT on DGT's translation workflow. More specifically, we looked at two different aspects, viz. productivity and perceived usefulness. In order to obtain results that are as close as possible to reality, we extracted the information from translations that were created in normal working conditions. We only minimally interfered in the normal translation workflow and asked the participants to enable/disable MT for half of the segments in each document. We collected data over a period of one month at the end of 2018 from 20 different translators of the French and Finnish language departments of DGT. All participants were volunteers, but given the high acceptance rate of MT usage by DGT translators [24], we assume that most results are also applicable on other language departments within DGT.

Extracting reliable time measurements from the data was more difficult than expected. SDL Trados Studio does not register when a segment is opened, but only when a segment is closed. Processing speed per segment was thus calculated as the difference between the closing of the current segment and the closing of the previous segment. This way of working posed problems for non-sequentially translated segments, and for this reason such segments were discarded in the analyses. As was done in other user studies [21,22] we also made use of two time thresholds. We used the same thresholds as defined by Federico et al. [21] and discarded segments with time measurements below 0.5 s per word and above 30 s per word.

For 1766 segments and a total of 34370 source words that were translated from scratch or based on MT, we extracted reliable time measurements. All users of NMT-FI and most users of SMT-FR were faster when using machine translation. The average speed gain of participating translators was 14% for NMT-FI and 12% for SMT-FR. This is much lower than the figures reported in other studies: Plitt and Masselot [27] reported that MT allowed translators to improve their throughput on average by 74%, Federico et al. [21] reports an average of 27% and Parra Escartín and Arcedillo [22] an average of 24%. It is of course difficult to compare the results of the different studies as the texts that were translated belonged to different domains (e.g., IT was included in all three above-mentioned studies) and the experimental conditions differed.

For a selected number of translators we had enough data to look at individual differences. We observed a large variation of individual processing speed among translators. The average from-scratch-translation speed varies between 5.01 s/token (approx. 720 words/hour) to 8.02 s/token (approx. 450 words/hour). With MT, the difference between translators is even larger and ranges from 3.96 s/token (approx. 910 words/hour) to 8.76 s/token (approx. 410 words/hour). We used the individual translation-from-scratch time measurements as a reference to compare speed gains or losses when using MT. For the translators who worked faster when using MT, relative speed gains ranged from 10% to 21%. Two French translators were slower when using SMT than when translating from scratch with relative speed losses of 2% and 36%. The results for the Finnish translators were more consistent, which might be attributed to the higher quality of the NMT engine. Some segments were considered good enough and left unchanged by translators (8% and 6% of machine-translated segments of the French and Finnish data set, respectively). Please note that this is only the case for shorter sentences, as the average sentence length of these segments is 9.20 for French and 11.44 words for Finnish. Even though these segments were not subject to any editing, they were not approved instantly, as the translators needed to check their accuracy. This residual amount of time can be considered incompressible, i.e., a minimum post-editing time assuming perfect machine translation quality. On average, the time translators spent on such segments was 34% of the time they would have spent translating the segment from scratch for NMT-FI (2.05 s/token or approx. 1756 words/hour), and 55% for SMT-FR (3.22 s/token or approx. 1118 words/hour).

It should be kept in mind that the figures we report capture only the effect of machine translation on a limited subset of the translators' tasks, namely translating those segments where machine translation can be put to good use. The segments we analysed represent 42–47% of all translated words. The net effect of machine translation on the overall translation workflow is thus considerably lower if all other translators' tasks (including translating segments with translation memory matches, self-revision, revision of other translations, etc.), on which machine translation has no effect, were considered.

As collecting reliable time measurements is not straightforward, we also examined whether there is a correlation between technical and temporal effort measurements, which if there is, would enable us to use technical effort as an approximation for temporal effort. We used word-based (HTER) as well as character-based metrics (CharacTER). On average, we observed higher technical effort (i.e., more edits) for NMT-FI than SMT-FR ($HTER_{fr} = 0.29, HTER_{fi} = 0.41; ChTER_{fr} = 0.23, ChTER_{fi} = 0.32$). For HTER, this observation can simply be attributed to the agglutinative nature of the Finnish language, which results in sentences with fewer words compared to French and more word form changes. However, this observation is also confirmed by CharacTER, a character-level metric, which suggest that the measurement of higher technical effort for NMT-FI compared to SMT-FR is not due to the agglutinative nature of the Finnish language.

The relationship between the technical and temporal effort in post-editing shows large variations across studies, with correlation values ranging from 0 (zero) to 0.524 [28–30]. In this study, we observed only weak correlations between total post-editing time (PET) and both technical effort measurements (HTER and CharacTER). The correlation coefficients were higher for Finnish than for French (Spearman correlation coefficient values PET/HTER of 0.327 for French and 0.401 for Finnish and Pearson correlation coefficient values PET/HTER of 0.157 for French and 0.305 for Finnish). In other words, the two technical effort measurements are not good indicators of the temporal effort involved in the post-editing task in this study.

At the end of the one-month logging period we set up a survey and asked the participants about their experiences with MT. Most translators think that they work faster or much faster when using MT. The perception of the translators is thus in line with the actual results. Reduction of typing effort is most frequently mentioned as reason why participants preferred working with MT, but also the psychological benefits of not having to start from scratch were often mentioned.

## 6. Conclusions

The study shows that on average, machine translation provides measurable benefits in real-life translation scenarios, and that these benefits are more consistent for NMT-FI than for SMT-FR. This complements research findings on the assessment of neural vs. statistical machine translation systems, which found that neural translation systems seem to provide more consistently useful output than statistical ones. The average speed gain we observed was 14% for NMT-FI and 12% for SMT-FR, which is much lower than the figures reported in other studies.

The main strength of this study is at the same time its main limitation: we worked with *real* translations, created in normal working conditions. Due to various reasons we could only extract reliable time measurements for a reduced set of segments and neglected other translation-related tasks such as project management, file engineering and reviewing.

This study was limited to two language pairs and studied the impact of MT on the translation workflow at DGT. In future work we want to extend this research by looking at more language pairs, different document types and compare the impact of MT on the workflows of different translation service providers.

**Author Contributions:** Conceptualization, L.M., D.P. and A.T.; methodology, L.M., D.P. and A.T.; software, A.T.; validation, L.M. and A.T.; formal analysis, A.T.; investigation, L.M., D.P. and A.T.; resources, L.M., D.P. and A.T.; data curation, L.M., D.P. and A.T.; writing—original draft preparation, L.M., D.P. and A.T.; writing—review and editing, L.M., D.P. and A.T.; visualization, A.T. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Acknowledgments:** We would like to thank the translators at the DGT, who willingly agreed to participate in this study and Baudouin Simon in particular.

**Conflicts of Interest:** The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

## Abbreviations

The following abbreviations are used in this manuscript:

| | |
|---|---|
| AVG | average |
| DGT | Directorate-General for Translation of the European Commission |
| FI | Finnish |
| FR | French |
| HTER | human-targeted translation edit rate |
| MT | machine translation |
| NMT | neural machine translation |
| NMT-FI | English–Finnish neural MT engine |
| PBSMT | phrase-based statistical machine translation |
| PET | post-editing time |
| RBMT | rule-based machine translation |
| SMT-FR | English–French phrase-based statistical MT engine |
| TMX | Translation Memory eXchange |

## References

1. Pigott, I.M. The importance of feedback from translators in the development of high-quality machine translation. In *Practical Experience of Machine Translation*; Lawson, V. Ed.; North-Holland Publishing Company: Amsterdam, The Netherlands; New York, NY, USA; Oxford, UK, 1982; pp. 61–74.
2. Wagner, E. Post-editing SYSTRAN, a challenge for Commission Translators. In *Terminologie et Traduction*; Commission des Communautés Européennes: Luxembourg, 1985; pp. 1–7.
3. Koehn, P.; Hoang, H.; Birch, A.; Callison-Burch, C.; Federico, M.; Bertoldi, N.; Cowan, B.; Shen, W.; Moran, C.; Zens, R.; et al. Moses: Open Source Toolkit for Statistical Machine Translation. In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Companion Volume Proceedings of the Demo and Poster Sessions, Prague, Czech Republic, 23–30 June 2007; pp. 177–180.
4. Drugan, J.; Strandvik, I.; Vuorinen, E. Translation quality, quality management and agency: Principles and practice in the European Union institutions. In *Translation Quality Assessment*; Moorkens, J., Castilho, S., Gaspari, F., Doherty, S., Eds.; Springer: Cham, Switzerland, 2018; pp. 39–68.
5. Krings, H.P. Repairing texts. In *Empirical Investigations of Machine Translation Post-Editing Processes*; Kent State University Press: Kent, Ohio, 2001.
6. Snover, M.; Dorr, B.; Schwartz, R.; Micciulla, L.; Makhoul, J. A study of translation edit rate with targeted human annotation. In Proceedings of the Association for Machine Translation in the Americas, Cambridge, MA, USA, 8–12 August 2006; pp. 223–231.
7. Lacruz, I.; Denkowski, M.; Lavie, A. Cognitive Demand and Cognitive Effort in Post-Editing. In Proceedings of the eleventh conference of the Association for Machine Translation in the Americas, Workshop on Post-editing Technology and Practice, Vancouver, BC, Canada, 22–26 October 2014; pp. 73–84.
8. Daems, J.; Vandepitte, S.; Hartsuiker, R.; Macken, L. Identifying the machine translation error types with the greatest impact on post-editing effort. *Front. Psychol.* **2017**, *8*, 15. [CrossRef] [PubMed]

9. Daems, J.; Macken, L. Interactive adaptive SMT versus interactive adaptive NMT: A user experience evaluation. *Mach. Transl.* **2019**, *33*, 117–134. [CrossRef]
10. Herbig, N.; Pal, S.; Vela, M.; Krüger, A.; van Genabith, J. Multi-modal indicators for estimating perceived cognitive load in post-editing of machine translation. *Mach. Transl.* **2019**, *33*, 91–115. [CrossRef]
11. Lesznyák, Á. Hungarian translators' perceptions of Neural Machine Translation in the European Commission. In *Proceedings of Machine Translation Summit XVII Volume 2: Translator, Project and User Tracks*; European Association for Machine Translation: Dublin, Ireland, 2019; pp. 16–22.
12. Aziz, W.; Castilho, S.; Specia, L. PET: A Tool for Post-editing and Assessing Machine Translation. In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey, 23–25 May 2012; European Language Resources Association (ELRA): Istanbul, Turkey, 2012; pp. 3982–3987.
13. Carl, M. Translog-II: A Program for Recording User Activity Data for Empirical Reading and Writing Research. In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), , Istanbul, Turkey, 23–25 May 2012; European Language Resources Association (ELRA): Istanbul, Turkey, 2012; pp. 4108–4112.
14. Alabau, V.; Bonk, R.; Buck, C.; Carl, M.; Casacuberta, F.; García-Martínez, M.; González, J.; Koehn, P.; Leiva, L.; Mesa-Lao, B.; et al. CASMACAT: An open source workbench for advanced computer aided translation. *Prague Bull. Math. Linguist.* **2013**, *100*, 101–112. [CrossRef]
15. O'Brien, S. Pauses as indicators of cognitive effort in post-editing machine translation output. *Across Lang. Cult.* **2006**, *7*, 1–21. [CrossRef]
16. Lacruz, I.; Shreve, G.M.; Angelone, E. Average pause ratio as an indicator of cognitive effort in post-editing: A case study. In Proceedings of the Tenth Conference of the Association for Machine Translation in the Americas, Workshop on Post-Editing Technology and Practice, San Diego, CA, USA, 28 October–1 November 2012; pp. 21–30.
17. Doherty, S.; O'Brien, S.; Carl, M. Eye tracking as an MT evaluation technique. *Mach. Transl.* **2010**, *24*, 1–13. [CrossRef]
18. Carl, M.; Dragsted, B.; Elming, J.; Hardt, D.; Jakobsen, A.L. The process of post-editing: A pilot study. *Cph. Stud. Lang.* **2011**, *41*, 131–142.
19. Daems, J.; Vandepitte, S.; Hartsuiker, R.; Macken, L. Translation methods and experience: A comparative analysis of human translation and post-editing with students and professional translators. *Meta J. Des Traducteurs/Meta Transl. J.* **2017**, *62*, 245–270. [CrossRef]
20. Läubli, S.; Fishel, M.; Massey, G.; Ehrensberger-Dow, M.; Volk, M. Assessing post-editing efficiency in a realistic translation environment. In Proceedings of the MT Summit XIV Workshop on Post-editing Technology and Practice, Nice, France, 2 September 2013; pp. 83–91.
21. Federico, M.; Cattelan, A.; Trombetti, M. Measuring user productivity in machine translation enhanced computer assisted translation. In Proceedings of the Tenth Conference of the Association for Machine Translation in the Americas (AMTA), San Diego, CA, USA, 28 October–1 November 2012; AMTA: Madison, WI, USA, 2012; pp. 44–56.
22. Parra Escartín, C.; Arcedillo, M. Machine translation evaluation made fuzzier: A study on post-editing productivity and evaluation metrics in commercial settings. In Proceedings of the MT Summit XV, Miami, FL, USA, 30 October–3 November 2015; pp. 131–144.
23. Cadwell, P.; O'Brien, S.; Teixeira, C.S. Resistance and accommodation: Factors for the (non-)adoption of machine translation among professional translators. *Perspectives* **2018**, *26*, 301–321. [CrossRef]
24. Rossi, C.; Chevrot, J.P. Uses and perceptions of Machine Translation at the European Commission. *J. Spec. Transl. (JoSTrans)* **2019**, *31*, 177–200.
25. Wang, W.; Peter, J.T.; Rosendahl, H.; Ney, H. Character: Translation edit rate on character level. In Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers, Berlin, Germany, 11–12 August 2016; pp. 505–510.
26. Schober, P.; Boer, C.; Schwarte, L.A. Correlation coefficients: Appropriate use and interpretation. *Anesth. Analg.* **2018**, *126*, 1763–1768. [CrossRef] [PubMed]
27. Plitt, M.; Masselot, F. A productivity test of statistical machine translation post-editing in a typical localisation context. *Prague Bull. Math. Linguist.* **2010**, *93*, 7–16. [CrossRef]
28. O'Brien, S. Towards predicting post-editing productivity. *Mach. Transl.* **2011**, *25*, 197. [CrossRef]

29. Gaspari, F.; Toral, A.; Naskar, S.K.; Groves, D.; Way, A. Perception vs reality: Measuring machine translation post-editing productivity. In Proceedings of the Third Workshop on Post-Editing Technology and Practice, Vancouver, BC, Canada, 22–26 October 2014.
30. Moorkens, J.; O'brien, S.; Da Silva, I.A.; de Lima Fonseca, N.B.; Alves, F. Correlations of perceived post-editing effort with measurements of actual effort. *Mach. Transl.* **2015**, *29*, 267–284. [CrossRef]