informatics

MDPI

Article

Translation Quality and Error Recognition in Professional Neural Machine Translation Post-Editing

Jennifer Vardaro *, Moritz Schaeffer and Silvia Hansen-Schirra

English Linguistics and Translation Studies, Johannes Gutenberg University, 76726 Mainz/Germersheim, Germany; mschae01@uni-mainz.de (M.S.); hansenss@uni-mainz.de (S.H.-S.)

* Correspondence: vardaro@uni-mainz.de

Received: 30 April 2019; Accepted: 11 September 2019; Published: 17 September 2019

Abstract: This study aims to analyse how translation experts from the German department of the European Commission's Directorate-General for Translation (DGT) identify and correct different error categories in neural machine translated texts (NMT) and their post-edited versions (NMTPE). The term *translation expert* encompasses *translator*, *post-editor* as well as *revisor*. Even though we focus on neural machine-translated segments, *translator* and *post-editor* are used synonymously because of the combined workflow using CAT-Tools as well as machine translation. Only the distinction between *post-editor*, which refers to a DGT translation expert correcting the neural machine translation output, and *revisor*, which refers to a DGT translation expert correcting the post-edited version of the neural machine translation output, is important and made clear whenever relevant. Using an automatic error annotation tool and the more fine-grained manual error annotation framework to identify characteristic error categories in the DGT texts, a corpus analysis revealed that quality assurance measures by post-editors and revisors of the DGT are most often necessary for lexical errors. More specifically, the corpus analysis showed that, if post-editors correct mistranslations, terminology or stylistic errors in an NMT sentence, revisors are likely to correct the same error type in the same post-edited sentence, suggesting that the DGT experts were being primed by the NMT output. Subsequently, we designed a controlled eye-tracking and key-logging experiment to compare participants' eye movements for test sentences containing the three identified error categories (mistranslations, terminology or stylistic errors) and for control sentences without errors. We examined the three error types' effect on early (first fixation durations, first pass durations) and late eye movement measures (e.g., total reading time and regression path durations). Linear mixed-effects regression models predict what kind of behaviour of the DGT experts is associated with the correction of different error types during the post-editing process.

Keywords: neural machine translation; post-editing; revision; error annotations; Hjerson; MQM; European Commission (DGT); eye-tracking; key-logging; post-editing effort

1. Background and Related Research

With "28 member states, 500 million citizens, 3 alphabets and 24 official languages" ([1], p. 1)—i.e., 552 possible language combinations—translation quality plays a key role in the European Union (EU) to bridge linguistic gaps and ensure successful communication between its citizens and institutions.

Approximately 1600 in-house translators make the European Commissions' Directorate- General for Translation (DGT) probably the world's largest translation service: In 2016 alone, the DGT received 73,000 translation requests, produced 2.2 million pages, and outsourced more than 650,000 pages to freelancers ([2], p. 43). With such a high workload, the recent paradigm shift from statistical to neural machine translations, and the strive to maintain extremely high-quality standards, the DGT is constantly adapting to new challenges, such as the creation of clear and consistent guidelines or

the development of their own machine translation systems. Statistical machine translation (internally referred to as MT@EC) has been available at the DGT since 2013 and is now coming to an end. The new neural system, eTranslation, was launched in November 2017 and is able to

> *"translate between any pair of the 24 official EU languages, as well as Icelandic and Norwegian (Bokmål): it can handle formatted documents and plain text; it translates multiple documents into multiple languages in "one go"; it accepts diverse input formats including XML and PDF; it retains formatting; and it provides specific output formats for computer-aided translation, i.e., TMX11 and XLIFF." ([1], p. 3)*

It is trained with the aligned source and target segments that have been produced over the years—amounting to approximately 1.2 billion training segments ([1], p. 3).

However, the usage of machine translation requires a different form of revision referred to as post-editing. Usually, machine translation users choose between three possibilities according to the purpose, life cycle, and (the size of) the target group of the text: (1) not correcting the MT output at all, (2) applying *light* corrections to ensure the understandability of the text or (3) applying a *full* post-editing which requires linguistic as well as stylistic changes to achieve a text of publishable quality that reads well and does not contain any errors. In the case of the DGT and its extremely high-quality standards, only *full* post-editing is applied and even the post-edits are revised by a second person. This is partly due to the DGT workflow which combines the use of translation memory systems (TMS) and machine translations (MT) in the same working environment. If a translator at the German department of the DGT decides to integrate MT in the translation process, target segment translations are displayed according to the preferences of translators and the prioritization of different TM segment match types. Context matches and 100% matches are usually displayed before MT, and for fuzzy matches, any desired threshold can be set (mostly 75% at the DGT), which means that fuzzy matches above said threshold will be displayed before MT, and vice versa. If there is no match at all, the segment either remains empty for translation from scratch, or the MT is blended in. In summary, the MT is not shown for the entire text simultaneously, but segment per segment and on demand. The translator is thus presented with MT segments in the same way 100% matches, etc., appear without having to leave the familiar working interface or perform extra steps to be able to use MT. This means that human translations and post-edits occur in one and the same text. For human translations, the workflow at the DGT generally foresees a revision by a second person, which has not changed since the arrival of integrated MT systems, which means that the post-edited segments in these translations are automatically reviewed by a second person. The whole DGT workflow complies with the requirements of various standards, e.g., the ISO standard for translation services [3] and the ISO standard for post-editing of machine translation [4].

The topic of (machine) translation quality in general has been researched from many different angles, e.g., in publications dealing with translation quality at the European Institutions, see [5–8], or with translation quality management, particularly at the DGT, e.g., [2]. Furthermore, there are more specialized guidelines for every language department. Translation quality evaluations can, e.g., be performed through manual or automatic error annotations. In the era of machine translation, more factors have entered the evaluation stage, such as automatic MT evaluation metrics or the search for possibilities to reduce post-editing effort involved in correcting MT.

However, the effect of neural machine translated text (NMT) suggestions on PE-effort is a relatively new research field which has not yet been investigated extensively. Applying linear mixed-effects regression models (LMERs, see below), [9] investigated the effect of MT error weights on PE effort indicators with the help of an eye-tracker device. They showed that average MT error weights are good predictors of six different PE effort indicators, namely average number of production units, average duration per word, average fixation duration, average number of fixations, pause ratio, and average pause ratio. They also found that "the different post-editing effort indicators are predicted by different MT error categories, with mistranslations, structural issues and word order issues being the most common categories" [9]. Participants were 10 translation students with no prior PE experience.

Texts were taken from different English newspaper articles and translated using the statistical MT system Google Translate (English into Dutch). They then expanded their study [10] using more fine-grained analyses to "verify whether all effort indicators are influenced by machine translation quality, and then identify the specific types of machine translation errors that have the greatest impact on each of the effort indicators" as well as both professional and student data. Text type and machine translation system remained the same, i.e., 13 professionals with 2–18 years of translation experience and 10 translation students at Masters level either post-edited or translated from scratch eight newspaper articles, 7–10 sentences long and covering various topics, which had been translated by the statistical Google Translate. Subsequently, two of the authors applied a two-step error annotation approach [11] to each translation, consisting of (1) acceptability problems such as grammar, syntax, style or register, and (2) of adequacy problems such as deletions, mistranslations or additions. This would correspond to the fluency and accuracy dichotomy used by MQM, which we apply in our study because it "provides a flexible framework for defining custom metrics for the assessment of translation quality" [12]. Contrary to our study, the two authors also assigned error weights ranging from 0 to 4 and calculated the average error weight per sentence. Only those error annotations on which both annotators agreed were analysed further. Results show that MT quality has an impact on both product and process PE-effort indicators but "not all process effort indicators seem to be influenced by the same machine translation error type".

In our study, we also investigate how the post-editing process is influenced by MT errors, but use non-weighted error categories and look at the PE-effort indicators total reading times, regression path durations, first pass durations, and first fixation durations in addition to the eye–key span (see below). We also expect to find that different error types influence the post-editing effort indicators differently. However, we do not assign random translators to random text types which had been translated by a random MT system, but instead focus on translation professionals from the DGT and use their in-house neural NMT system eTranslation which they would normally use when post-editing a text. The stimuli in our experiments are also as authentic as they can be, i.e., error words taken from real-life post-editing scenarios at the DGT. Only the text around these error words was manipulated (as little as possible) to suit the requirements of a controlled experiment. A correct version of these error words was provided by the revisions conducted at the DGT. The manipulation and the correct version of the errors were necessary in order to (1) isolate eye movements related to a single error type from other eye movements and (2) to be able to compare the eye movement behaviour for an erroneous word with that of a correct word (see Section 2.3.2 Study Design). The authors of [10] cannot tie their results to specific error words that cause an increase in PE effort because they worked on a sentence level and their sentences often contained more than one error of various categories, which makes it difficult to isolate single processes. Furthermore, since they did not manipulate their texts, they can only study the effect of the error weights and not the effect of errors compared to the effect of correct versions.

Bentivogli et al. [13,14] conducted a product-based study reporting reduced PE effort for NMT compared to SMT, however, as [15] pointed out, a product-based analysis of PE changes in MT alone is an insufficient measure of the actual PE effort involved. Lacruz [16], e.g., also provided evidence that not only the number of errors but also the error type plays an important role in determining the actual PE-effort involved, because some errors seem to be more demanding than others. Recent studies observed that NMT reduces some error types more than others e.g., [17–19], some of them using process measures such as keystrokes or pauses. However, none of them has so far applied eye-tracking measures such as average fixation counts and durations, which, according to [20], correlate strongly with technical and temporal post-editing effort. Koponen [21] collected data with an eye-tracker but the quality of the recordings was insufficient for the mapping of eye movements to single words.

Our study draws much of its methodological framework from the eye-tracking study of revision processes conducted by [22], who were among the first to compare the reading behaviour of translation students with that of professional translators by analysing their eye movements when revising human translations. They presented a total of 748 errors (six categories adapted from Mertin's [23] taxonomy),

inserted in newspaper articles taken from the CRITT-TPR Database [24] to 23 professionals and to 15 students to investigate the effects these errors had on eye movement behaviour. Results show that professionals revise translations more efficiently than students. Although they corrected significantly more errors, all eye movement measures showed that professionals were faster. One of the reasons for this effect is that professionals seem to prioritize their search for errors in that they only refer to the source text when necessary, while students do so for errors that could be corrected by reading only the target text. More precisely, students tend to spend more time reading the source text irrespective of the error category, while professionals only spend time reading the source text significantly longer for sense errors. Furthermore, professionals tend to re-read previous text in the target text when they encounter sense or coherence errors, while students tend to do so when they encounter orthography errors, even though context seems to be rather irrelevant for the correction of this error type. All in all, professional translators seem to evince a more strategic and thus more efficient reading behaviour. Orthography errors are recognized early (first fixation duration and first pass duration) by students and professionals. However, sense, consistency and coherence errors only had an effect on late eye movement measures. This pattern of results could suggest that sub-lexical processes are engaged in the early recognition of orthographic errors, while sense, consistency and coherence errors only appear late, because they require integration of the lexical information in the broader context constituted by source and target aspects.

A drawback of the [22] study, as they point out, is that error types and error frequencies should have been identified beforehand in a suitable corpus, which is why we conducted a corpus analysis before designing the eye-tracking study. Furthermore, instead of professional revision processes for human translations, we analysed the reading behaviour of professional translators for neural machine translation post-editing (NMTPE). Similar to the aim of the [22] study, we try to address the following research questions: Does the error type have an effect on eye movements, i.e., on how early or late an error is detected? Which error types are recognized earlier than others? Are there any eye movement patterns which can be linked to error recognition and correction? By trying to answer these questions, we hope to shed light on the cognitive processes regarding error recognition and post-editing effort of machine translation. In the field of natural language processing (NLP), a recent development has gained more and more attention, namely Quality Estimation (QE). The aim here is to predict the quality of machine translations to give an estimate on how good or bad the translation is. The main difference from standard evaluation measures like BLEU is that QE systems estimate the quality without access to (human) reference translations [25].

The foundation of QE systems are supervised machine learning tasks using different algorithms. Models are created from translation examples which are described through various features and annotated for quality [25]. Or as [26] put it:

> *The common approach to automatic translation quality estimation is to transform the problem into a supervised regression or classification task for sentence-level scoring and word-level labelling respectively.*

In this context, [27], e.g., propose a "novel supervised regression model for the segment-level MTE [automatic machine translation evaluation] based on universal sentence embeddings". Other methods, where positive results have been reported, include improving post-editing efficiency by filtering out low-quality segments which would require more effort or time to be corrected than translating from scratch [28,29], or highlighting the parts that need revision [30]. If these methods achieve a relatively high accuracy, a reduction of cognitive effort for the post-editor is entailed.

Quality estimation models can be regarded as a parallel to the human process of error recognition. The models 'learn' to extract and analyse (linguistic) features from texts, to take into account additional resources such as translation tables, and to use this acquired 'knowledge' to flag any (mostly linguistic) breaches in new machine translations. In consequence, they are, e.g., able to estimate whether it is necessary to transform a breached text into an acceptable text. This is exactly what a human translator would also do: perceive and analyse words by transforming them into linguistic representations,

compare these representations with each of the languages in question, and detect where matches or errors are. Some work has been published on document-level quality estimation, e.g., [31,32], but human translators still have a big advantage when it comes to taking extra-linguistic and cultural context into account. In a nutshell, QE models can be trained, i.e., they regularly improve or even 'learn' new features. Human translators are by today's standards far ahead, and also thought to develop a higher degree of expertise the more they practice. The mechanisms which develop in humans with exposure to the task in question can thus be seen as equivalent of what QE approaches attempt to model.

In this study, we attempt to model the cognitive processes in humans, which may be argued to be equivalent to the QE models developed in NLP. We use linear mixed-effects regression models (LMERs) and try to model the post-editing effort in humans on the basis of eye movement data. Note that we are only reporting initial results. We also recorded data from a group of students to compare their eye-movement and key-logging behaviour to that of the DGT professionals in a follow-up study. This will allow us to model the effect of translation expertise on error recognition processes more precisely.

2. Experimental Setup and Methods

2.1. Corpus Analysis

We used authentic DGT texts covering various topics to create a parallel corpus consisting of 24 English source texts (902 segments, 17,925 tokens), 24 German NMT outputs (902 segments, 15,347 tokens), 24 German NMTPE (902 segments, 15,396 tokens), and 24 German revisions (REV) of the NMTPE (902 segments, 15,319 tokens). More precisely, the corpus was compiled using the translation origin given by SDL Trados Studio, the current CAT-Tool of choice at the DGT. All translation units labelled with *Automatic Translation (AT)*, and only those, were extracted from the SDL packages that we were supplied with by the DGT and collected in an excel sheet. Thus, the term *texts*, in this case, refers to those parts of the original texts which had been processed automatically.

Post-editors and revisors of the texts in our corpus were all professionals employed by the DGT. It is not known to us who post-edited and who revised how many of the collected texts. However, it becomes clear from the workflow that one person can have post-edited more than one text and/or revised more than one text, but it is never the same person who post-edited *and* revised one and the same text. The changes these professionals performed were defined as errors to be analysed in the corpus analysis, i.e., any errors they may have missed are not included in this study. Note that not all performed changes are necessarily related to actual errors, which is why the MQM framework for manual error annotations uses the term *issues* for all categories. Here, we understand *error* as an umbrella term for *all* changes performed by the DGT experts, i.e., error refers to potential issues as well as to actual errors in a category. Future research involves dividing the MQM issues we used for our eye-tracking-study into preferential and essential changes to see if this yields any significant changes in the results presented in this paper. In a first step, we annotated the errors in our DGT corpus automatically using Hjerson [33], comparing NMT with NMTPE and NMTPE with REV. Hjerson identifies five error categories as defined by Vilar et al. [34]: extra word (ext), missing word (mis), inflection (infl), reordering (reord), and lexical error (lex). Upon inspection of the automatically annotated errors, it became obvious that the largest error category (lexical errors) contained several different types of errors, which is why we decided to perform a more fine-grained manual error annotation according to the MQM framework in a second step. Since extra words are difficult to distinguish from lexical errors, we decided to also annotate them manually to better understand what kind of errors were subsumed under this category. The MQM framework encompasses a great variety of error categories which are not supposed to be analysed all at once but should rather be tailored to specific use cases. Sample annotations of some of the sentences were performed using the MQM decision tree (http://www.qt21.eu/downloads/fullDecisionTreeComplete.pdf [Last openend: 27.04.2019]) to identify relevant error types within the lexical category. In the course of the annotation,

categories were constantly readjusted to resemble the error prevalence of the DGT corpus, although we tried to avoid having too many subclasses and thus too small datasets for the statistical analyses.

2.2. Corpus Analysis Results

The corpus analysis and its results have already been described in [35]. Here, we will outline relevant findings for a better understanding and possible reconstruction of the eye-tracking experiment.

2.2.1. Automatic Error Annotation with Hjerson

Hjerson automatically identifies words which contribute to the Word Error Rate (WER), the position-independent error rate in the reference (RPER) and the position-independent error rate in the hypothesis (HPER) [36]. The five error categories are based on [34] and defined as follows:

1. *inflectional error—a word whose full form is marked as RPER/HPER error but the base forms are the same.*
2. *reordering error—a word which occurs both in the reference and in the hypothesis is thus not contributing to RPER or HPER but is marked as a WER error.*
3. *Missing word—a word which occurs as deletion in WER errors and at the same time occurs as RPER error without sharing the base form with any hypothesis error.*
4. *extra word—a word which occurs as insertion in WER errors and at the same time occurs as HPER error without sharing the base form with any reference error.*
5. *incorrect lexical choice—a word which belongs neither to inflectional errors nor to missing or extra words is considered as lexical error [33].*

Figure 1 illustrates the error distribution (raw total error counts) in NMT and NMTPE after the automatic annotation with *Hjerson:*

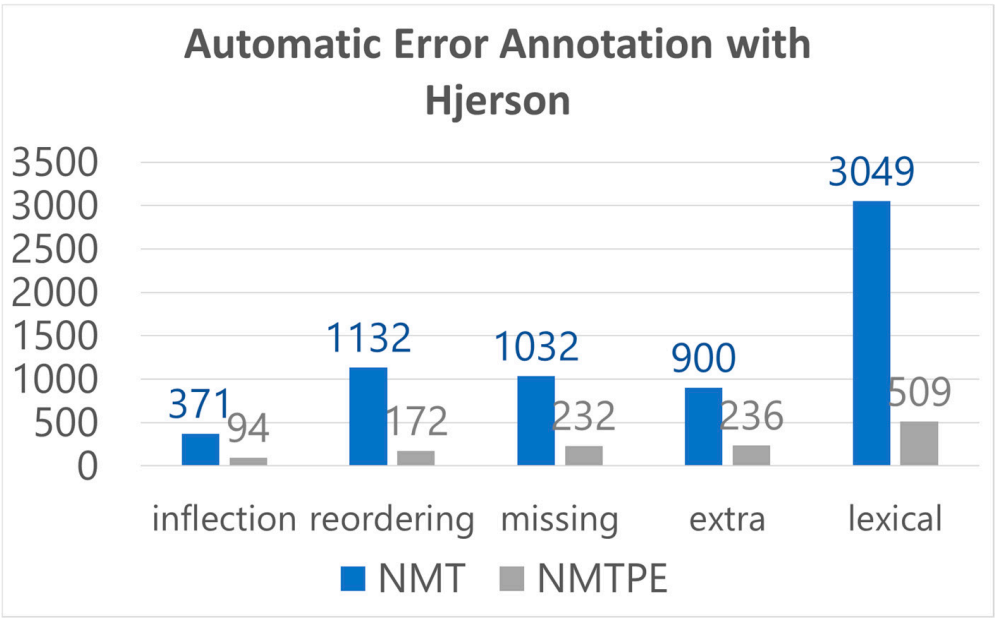

Figure 1. Raw Error Count after automatic annotation with Hjerson.

Lexical error (47.02%) was the most frequent NMT error category, followed by reordering (17.46%), missing word (15.29%), extra word (13.88%), and inflection (5.72%). The most frequent error category in NMTPE was again lexical error (40.95%), followed by extra word (18.99%), missing word (18.66%), reordering (13.84%), and inflection (7.56%). In the next step, we decided to manually annotate the most prominent error category, lexical errors, to gain a better understanding of its subclasses, as well as the category extra words, because they are difficult to distinguish from lexical errors.

2.2.2. Manual Error Annotation According to the MQM Framework

There are many metrics for the purpose of manual error annotations, e.g., as mentioned before. by Mertin [23] and Vilar [34], or the more recent SCATE taxonomy by Tezcan [37]. We opted for the MQM framework, because they looked at many already existing metrics and tried to combine most of their categories into a single taxonomy, i.e., they "created a master listing of all of the types of issues checked in different metrics" [38], with which one could represent most of the other taxonomies as well. This facilitates the usage of a consistent error annotation terminology and thus the comparability of studies.

Table 1 shows the six final MQM issues we used for the manual annotation of lexical errors and extra words in NMT and NMTPE:

Table 1. MQM issues, their definitions and Directorate-General for Translation (DGT) examples of manual error annotation.

MQM Issue	MQM Definition
Mistranslation (*Mistr*)	The target content does not accurately represent the source content.
DGT Example: • AT: **Interest**, scope, [. . .] and purpose of Binding Valuation Information • NMT: **Zinsen**, Umfang, [. . .] und Zweck der verbindlichen Bewertungsinformationen • NMTPE: **Interesse**, Umfang, [. . .] und Zweck der verbindlichen Wertermittlungsauskünfte • REV: **Relevanz**, Umfang, Funktionen und Zweck verbindlicher Zollwertauskünfte	
Terminology (*Term*)	A term (domain-specific word) is translated with a term other than the one expected for the domain or otherwise specified.
DGT Example: • AT: Interest, scope, functions and purpose of **Binding Valuation Information** • NMT: Zinsen, Umfang, Aufgaben und Zweck der verbindlichen **Bewertungsinformationen** • NMTPE: Interesse, Umfang, Funktionen und Zweck der verbindlichen **Wertermittlungsauskünfte** • REV: Relevanz, Umfang, Funktionen und Zweck verbindlicher **Zollwertauskünfte**	
Unidiomatic (*Phra*)	The content is grammatical, but not idiomatic.
DGT Example: • AT: If you **hire** crewmembers to work on ships . . . • NMT: Wenn Sie Besatzungsmitglieder **gemietet** haben, . . . • NMTPE: Wenn Sie Besatzungsmitglieder **eingestellt** haben, . . . • REV: Wenn Sie Besatzungsmitglieder **eingestellt** haben, . . .	
Register (*Reg*)	The text uses a level of formality higher or lower than required by the specifications or general language conventions.
DGT Example: • AT: We have to [. . .] counter the **causes** of climate change. • NMT: Wir müssen [. . .] die **Ursachen** des Klimawandels [. . .] bekämpfen. • NMTPE: Wir müssen [. . .] die **Auslöser** des Klimawandels [. . .] bekämpfen. • REV: Wir müssen [. . .] die **Auslöser** des Klimawandels [. . .] bekämpfen.	
Spelling (*Spel*)	Issues related to spelling of words
DGT Example: • AT: Informal meeting in **Sibu** • NMT: Informelles Treffen in **Sibu** • NMTPE: Informelles Treffen in **Sibiu** • REV: Informelles Treffen in **Sibiu**	
Function words (*Func*)	A function word (e.g., a preposition, "helping verb", article, determiner) is used incorrectly.
DGT Example: AT: **On** the proposal of President Juncker, . . . NMT: **Zum** Vorschlag von Präsident Juncker . . . NMTPE: **Auf** Vorschlag von Präsident Juncker . . . REV: **Auf** Vorschlag von Präsident Juncker . . .	

The *MQM* category definition for *register* (*Reg*) was slightly adapted and encompasses all possible language variants, even if they were classified as belonging to the same register. Therefore, this category will also be referred to as *stylistic changes* in the remainder of the paper.

Since automatic error annotation has its flaws, the subsequent manual error annotation was performed to clean up erroneous annotations and to gain more insight into the subclasses that make up the lexical and extra category.

In total, 3325 errors were annotated in NMT (see Figure 2). In total, 1962 (59.01%) of all lexical NMT errors fall into the register category, followed by 655 function words (19.7%), 429 (12.9%) mistranslations, and 201 terminology errors (6.05%). The remaining two categories are negligible. In total, 624 annotations were considered erroneous, which is mainly due to the fact that Hjerson classifies different punctuation tokens as lexical errors since there is no separate error category for punctuation issues.

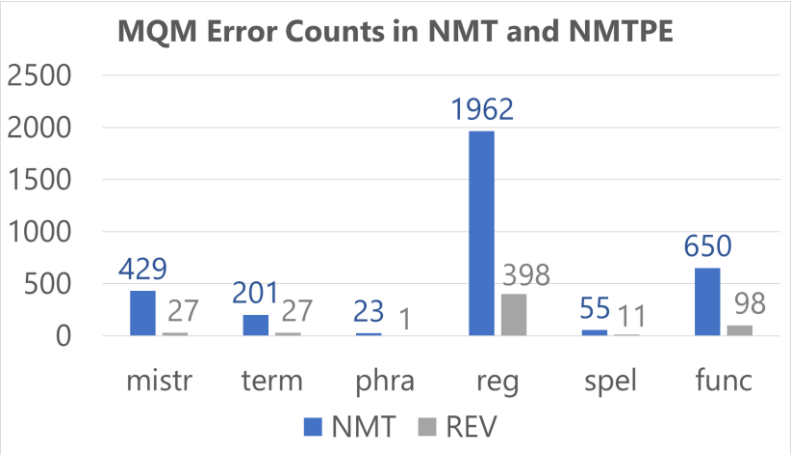

Figure 2. Error distribution after manual MQM annotation of NMT.

In total, 562 errors were annotated in NMTPE (see Figure 2). The relatively small number of NMTPE annotations limits the generalizability of this study to a certain extent but conclusions for our specific use case at the DGT can still be drawn. In total, 398 (70.82%) of all lexical NMTPE errors fall into the register category and 98 (17.44%) errors are function words, followed by 27 mistranslations (4.8%) and 27 terminology errors (4.8%). The remaining two categories are negligible. In total, 183 annotations were considered erroneous, again mainly due to punctuation issues which we did not focus on in this study.

Results suggest that post-editors either missed some of the errors or, if they corrected them, did not perform the right changes, so revisors had to interfere again. To be able to cast a final judgement on this, however, it needs to be further clarified whether the changes the revisors performed (1) affected the same words that the post-editors had already corrected and (2) were of preferential or rather essential nature. The fact that many stylistic changes were performed at both the post-editing and the revision stages might seem surprising at first glance. However, given the extremely high quality standards of the DGT and the fact that they treat post-editing exactly the same as human translations during their workflow, it was to be expected. They are not trained to be highly efficient post-editors, where the best quality is not always necessary and stylistic changes are usually not supposed to be inserted, but rather used to putting a lot of thought and time into their corrections to create a translation that is as perfect as possible, which leads to possible over-editing. Another factor that could play a role is priming effects, evoked by external stimuli which may affect subsequent responses by activating mental constructs without conscious realization [39]. Priming effects of machine translation outputs could be an explanation for the similar prevalence of error categories in NMT and NMTPE: the machine translation system produces an error, the post-editors are primed by this type of error and thus sometimes unable to find the right solution, so that, ultimately, the revisor has to correct the same error type again. Priming effects caused by MT systems have been reported before, e.g., by Green [40], also covering different aspects of translation such as the use of terminology in MT and PE [41], MT markers in PE [42] or the evidence of *post-editese* [43].

2.2.3. Effect of NMT Error Types on Errors in NMTPE

It proved to be rather difficult to distinguish between terminology errors and mistranslations during the annotation process, since the annotations were not performed by DGT professionals but instead by an independent lecturer from our translation faculty, who cannot be considered an expert in EU terminology. Therefore, we decided to merge terminology errors and mistranslations into a single category (*Misterm*). We used simple linear regressions with the same error category as the predictor and dependent variable to test to what extent error types in the raw NMT segments could predict the number of errors of the same error type still present in the post-edited segments. In other words, we predicted the effect of NMT errors on NMTPE errors per sentence. Only *Misterm* errors and stylistic changes (*Reg*) in the NMT sentences had a significant effect on the same error types in NMTPE sentences, as illustrated in Figures 3 and 4.

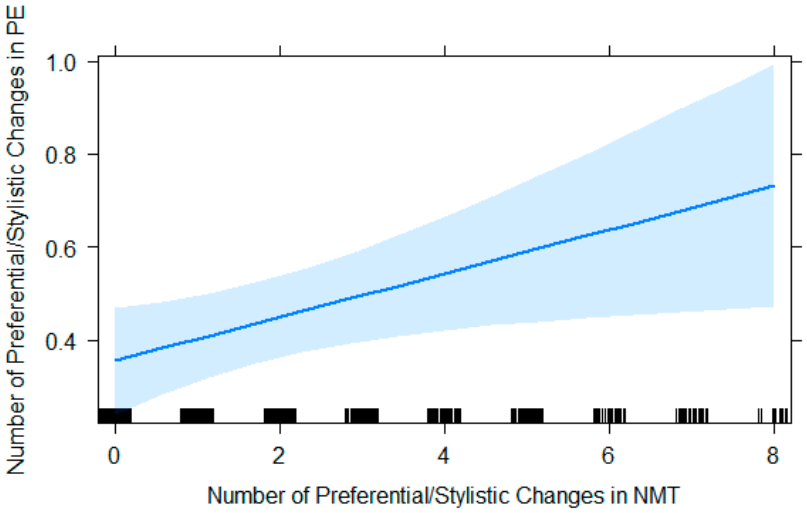

Figure 3. Effect of register errors in NMT on register errors in NMTPE.

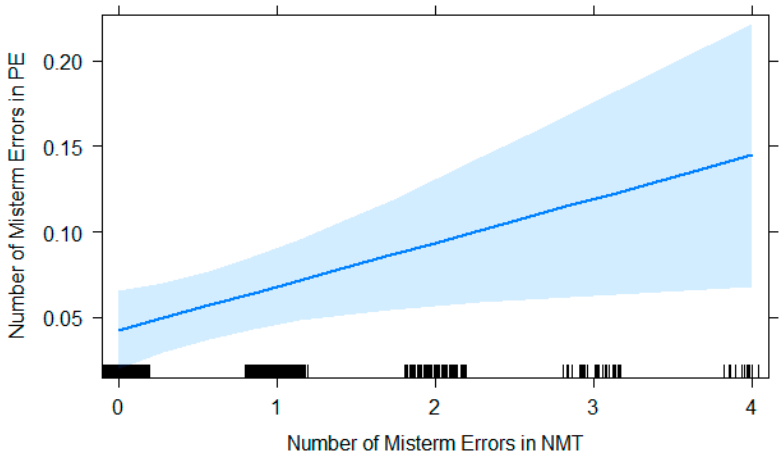

Figure 4. Effect of *Misterm* errors in NMT on NMTPE *Misterm* errors.

The effect of stylistic changes in NMT sentences on stylistic changes in NMTPE sentences was positive and significant ($\beta = 0.05$; $SE = 0.02$; $t = 2.31$; $p < 0.05$). For every register error in the NMT, there were 0.05 register errors per sentence in the NMTPE.

Figure 5 illustrates the effect of mistranslations and terminology errors in NMT on mistranslations and terminology errors in NMTPE per sentence, which was positive and significant ($\beta = 0.03$; $SE = 0.01$; $t = 2.3$; $p < 0.05$). For every mistranslation or terminology error in NMT sentences, there were 0.03 errors in NMTPE sentences.

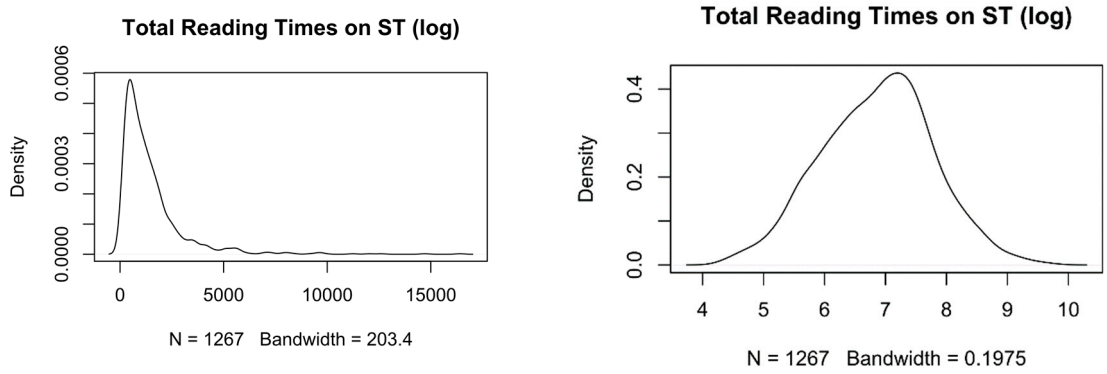

Figure 5. Density total reading times on ST (**left**) and the same density log-transformed (**right**).

The two small yet significant effects of these error categories in the NMT sentences shining through in the NMTPE sentences indicate that post-editors at the DGT are, as explained earlier, indeed being primed by the NMT output to a certain extent, particularly by mistranslations/terminology errors and stylistic/register errors. Note that our analyses were conducted at the sentence level. Future analyses will also examine to what extent (1) the same words in these same sentences were affected and thus corrected twice, (2) revisors detected errors that the post-editors had overlooked, and (3) revisors corrected errors that the post-editors introduced. Another interesting factor would be to look at cross-categorial relationships of errors using logistic regression models instead of linear regression models to see whether the prevalence of one error category in the NMT has an effect on different error categories in the NMTPE.

Subsequently, we analysed the most prominent error categories in the DGT corpus, i.e., mistranslations/terminology errors, stylistic changes and function words in an eye-tracking and key-logging study to gather post-editing process data and be able to look at these error categories from a different angle.

2.3. Eye-Tracking and Key-Logging Experiment

2.3.1. Participants

Conducting experimental research with highly proficient participants, texts they are used to working with, and corresponding translations as well as revisions which had been created in real-life scenarios is not only very interesting but also contributes to the ecological validity of our experiment. Ecological validity can be described as the degree to which the results of controlled experiments are related to those obtained in naturalistic environments [44]. Our study took place at the German Department of the DGT in Brussels and in Luxemburg. Thirty professional translators from this department took part in the study. Demographic and Language History data, however, was only available from 27 of these. All but two were L1 speakers of German; one had English as L1 and one Luxembourgian. They were not excluded from our study because the DGT considers them proficient enough to work for their German department, i.e., *into* German. Table 2 provides more detailed information on the participants and their language history.

In other words, participants in this study learned their languages late; they are rather balanced in their language use; and they mostly use their L2 with friends and/or colleagues. All participants stated that their working target language was German, but they regularly translate from between 1 and 6 different source languages. We can summarize that these participants were truly multilingual and highly proficient. The degree of bilingualism is routinely reported in the literature because it has an effect on behaviour during translation [45].

Considering the expertise of participants, we expect that they recognize errors rather early, even though there might be differences across categories. The authors of [22] found differences between

early recognized categories such as orthography, where a very small part of the target text already contains sufficient information, and categories that only had an effect on late measures, such as sense, coherence or cohesion, where the source text and/or big chunks of the target text are needed more often.

Table 2. Demographic and language history data of eye-tracking study participants.

	Mean (SD)	Additional Information
Demographic Data		
Age in years	45 (9.9)	
Gender (F:M)		(18:9)
Language Background		
L2		English (67%), French (19%), German (7%), Spanish (4%), Italian (4%)
L3		French (63%)
L4		81% had an L4: Spanish (27%), Italian (23%), English (18%), French (9%), Russian (9%) or either Luxembourgish, Dutch or Portuguese (5% each)
L5		52% had an L5 (French, Italian, Lithuanian, Dutch, Portuguese, Czech, Russian, Swedish, Spanish)
School subjects taught in L2		11% were regularly taught subjects such as biology or history in their L2
Learning environment L2		85% learned their L2 in a formal setting.
Learning environment L3		81% learned L3 in a formal setting
Starting age learning L2	10.2 (2.9)	
Starting age learning L3	14.6 (5.7)	
Years of study L2	12.1 (7.8)	
Language Use		
Speaking L2 (setting)		26% speak their L2 regularly with their families, or relatives 19%, most speak their L2 with their friends (74%) or colleagues (85%).
Speaking L2 (amount of time)		33% speak in their L2 between 1 and 15 h a day, 26% speak in their L2 between 1 and 15 h a week 40% speak in their L2 between 1 and 15 h a month
Reading L1 (hours per week)	16.3 (12.1)	
Consumption of audio-visual L1 material (hours per week)	5.6 (5.1)	
Reading L2 (hours per week)	15.6 (14.7)	
Consumption of audio-visual L2 material (hours per week)	4.1 (5.6)	
Language Competence (self-rated)		
General knowledge of L2	82.7 (10.7)	On a scale from 1 (very poor) to 100 (very good)
Active knowledge of L3	85.4 (8.3)	On a scale from 1 (very poor) to 100 (very good)
Passive knowledge of L3	82.5 (14.0)	On a scale from 1 (very poor) to 100 (very good)
Ability to translate from L2 into L1	92.7 (7.2)	On a scale from 1 (very poor) to 100 (very good)
Ability to translate from L1 into L2	72.5 (15.6)	On a scale from 1 (very poor) to 100 (very good)
Translated hours per week from L2 into L1	21.7 (16.2)	

2.3.2. Study Design

Broadly, two error type groups were used in the study by [22]: fluency errors such as orthography, grammar, consistency and coherence, on the one hand, and accuracy errors (sense and omission) on the other. Within the fluency group, a further distinction can be drawn: orthography errors involve sub-lexical aspects while omission, consistency and coherence errors require contextual integration for error recognition to occur. We decided to use the most frequent error categories in the corpus for our eye-tracking study, namely register errors, mistranslations, terminology errors and function words (for examples, see Table 2). According to the MQM framework, mistranslations and terminology errors can be grouped together as accuracy errors, whereas phraseology, register, spelling and function words are fluency indicators. However, the strict boundaries of this dichotomy may become blurred once we divide the error categories into essential and preferential changes.

To create the stimuli for the eye-tracking study, we manipulated texts consisting of source, test, control, and filler sentences as follows: We extracted suitable sentences from the NMT subcorpus, which contained either a *Misterm*, *Reg* or *Func* error and not too many other errors, if any. Only one error of one of the three error types was left in every sentence (in the middle region of the sentences, i.e., the error was never the first nor the last word and most often situated close to the middle of the sentence) and all other errors were corrected according to their final version, i.e., the revision. These sentences constitute the test sentences of our eye-tracking experiment. The decision to use manipulated sentences containing only one error of a single category was a matter of control and statistical validity, which can be justified by the fact that, this way, processes which affect only this particular error can be analysed in isolation from other processes that would occur if the sentences contained more than one error (category). We then prepared a second version of these sentences, the control sentences, by correcting the one remaining error according to the final version of the DGT, i.e., not all sentences in the experiment contained an error. Participants saw either the test or the control sentence, never both versions of the same sentence. Critical error words and their corrected versions did not differ significantly in terms of word frequency (*Misterm*: $t = -1.54$, $p < 0.13$; *Reg*: $t = -0.59$, $p < 0.56$; *Func*: $t = -0.81$, $p < 0.42$), number of cognates (*Misterm*: $t = 0$, $p < 1$; *Reg*: $t = 1.02$, $p < 0.31$; *Func*: $t = -0.47$, $p < 0.64$) or word length (*Misterm*: $t = 0.55$; $p < 0.58$, *Reg*: 0.42 $p < 0.68$; *Func*: $t = -0.96$, $p < 0.63$), to reduce the effect of potential irregularities, such as infrequent/long words or cognates, on participants to a minimum, so that the results of the behaviour analysis would not be distorted.

In a last step, we included some filler sentences which contained more than one error and different error categories to make the texts more realistic and to achieve a more balanced and non-systematic error distribution to prevent over- or under-editing by the participants.

The following examples in Table 3 illustrate the structure of the stimuli in our experiment:

The full text corpus used for the eye-tracking study is made available in the Supplementary Materials. Note that no test sentences containing extra words were included because, for it to be an extra word, there must be a missing word in the source and/or control sentence, which means that we would not have any source or control tokens to compare the test token with because eye-tracking data on missing words cannot be measured. In future studies, it would be interesting to not only analyse the post-editing behaviour but also the revision behaviour of the DGT professionals and compare it with the post-editing data obtained for this study. Investigating the two tasks in the same study was not possible since our experiment already took two hours for each participant and another study would have been out of scope.

In the study by [22], grammar errors were recognized early, and *Func* errors in the present study are grammatical in nature. *Misterm* errors in our study are inappropriate in context and should therefore require comparison with the source text—similarly to the sense errors in the [22] study, which is why we expected to observe similar effects to those regarding grammar and sense errors in [22]. Because of the subtle semantic differences between correct and erroneous register tokens, we expected that recognising register errors would show up only or mainly in late eye movement measures. In addition, given that register errors are, by definition, dependent on context, we expected

that they would often not be recognized as errors in the first place and if they were recognized, we hypothesized that this would, again, be a rather slow process.

Table 3. Example sentences for the stimuli structure in the eye-tracking experiment.

Func	Source Sentence	Martin Selmayr has been appointed today as the European Commission's new Secretary-General **on** the proposal of President Juncker.
	Test Sentence	Martin Selmayr wurde heute **zum [to]** Vorschlag von Präsident Juncker zum neuen Generalsekretär der Europäischen Kommission ernannt.
	Control Sentence	Martin Selmayr wurde heute **auf [on]** Vorschlag von Präsident Juncker zum neuen Generalsekretär der Europäischen Kommission ernannt.
Reg	Source Sentence	Since its set up in 2007, the European Chemicals Agency (ECHA) has a key role in the **implementation** of all the REACH processes.
	Test Sentence	Seit ihrer Gründung im Jahr 2007 spielt die Europäische Chemikalienagentur (ECHA) eine Schlüsselrolle bei der **Umsetzung [implementation]** aller REACH-Verfahren.
	Control Sentence	Seit ihrer Gründung im Jahr 2007 spielt die Europäische Chemikalienagentur (ECHA) eine Schlüsselrolle bei der **Durchführung [implementation]** aller REACH-Verfahren.
Misterm	Source Sentence	(7.) There have to be breaks during the working day.
	Test Sentence	(7.) Es muss **Brüche [fractures]** während des Arbeitstages geben.
	Control Sentence	(7.) Es muss **Pausen [pauses]** während des Arbeitstages geben.

Note that, due to the DGT workflow which combines TMs and MT, a manipulation of coherent texts beyond the segment level was impossible, which is why we opted for the manipulation of single sentences, which often stem from one and the same or a very similar DGT text but do not form a coherent text in themselves. We tried to identify sentences which are understandable without further context and grouped them together in blocks, hereinafter referred to as *texts*. This was mainly due to practical reasons: After each of these texts, the eye-tracker was recalibrated to ensure that recordings were not becoming inaccurate after a while, also because this way, participants had the chance to take a break between the recordings of the texts if and whenever needed. In total, 11 texts—each consisting of 9 single sentences—to 30 professional translators from the German Department of the DGT, i.e., 2970 critical and control tokens in a sentence context were presented. The order of texts was mixed in a Latin square.

All participants gave their informed consent for inclusion and publication before we started the recordings. Furthermore, we made sure that every participant was aware that they were supposed to post-edit single sentences grouped together for practical reasons, and not coherent texts. Participants were also told that not all sentences contained an error. Apart from that, participants did not know anything about the error structure and order of the sentences. However, they were presented with information on:

1. the user interface and functioning of the devices and tools needed for this study (see below)
2. the recommended (though not obligatory) time frame (7–9 min per text) to avoid very long and very short sessions
3. the engine used to produce the NMT (DGT in-house system: eTranslation)
4. the fact that they would not have access to in-house or external resources to conduct research
5. expected results according to the extremely high-quality standards of the in-house DGT guidelines (full post-editing, publishable quality)

The sessions were recorded using the non-invasive eye-tracking device SMI RED250Mobile (250 Hz) and the eye-tracking and key-logging tool Translog-II [46].

3. Data Analysis

The data obtained (xml files) was annotated with word and segment alignments of source and target texts in the YAWAT tool [47]. The statistical analysis with linear mixed-effect regression models (LMER) was carried out in *R* [48], using the package *lme4* [49]. This type of statistical analysis was chosen over more traditional approaches such as ANOVAs because regression models are more flexible, handle unbalanced data well, and differences between each token and participant can be accounted for by including participant and token as random effects in the regression model. Both fixed effects, i.e., predictors, and random effects can be included in the same model. The package *lmerTest* [50] was used to calculate standard errors, effect sizes, and significance values. The effects of the models were visualized in plots for a better interpretation of each model by applying the *effects* package [51].

There were four possibilities for the participants to interact with each of the tokens presented in our experiment:

1. They were presented with an error token (Err) and corrected the error. hereafter referred to as True Positives (TP)
2. They were presented with an error token and did not correct the error. hereafter referred to as False Positives (FP)
3. They were presented with a correct token (Corr) and did not change the token. hereafter referred to as True Negatives (TN)
4. They were presented with a correct token and changed the token. hereafter referred to as False Negatives (FN)

The errors left uncorrected (FP) were assumed to have remained unrecognized by the participants, or else they would have changed these error tokens. Since the aim of this study was to analyse the behaviour of DGT professionals when recognizing and correcting errors, we excluded FP, i.e., error tokens which were not modified (under-revision) and FN, i.e., correct tokens that were modified (hypercorrections) from the statistical analysis to be able to model the effect of recognized and corrected errors on eye movements correctly. Critical tokens are thus corrected errors, i.e., TP, while TN are correct control tokens in the final target sentences which remained untouched.

In the following, eye movement behaviour for error tokens in the test sentences is compared to the behaviour for correct tokens in the control sentences. This is included in the models as the predictor *Corrected*, which has two levels, *Correct (TN)* and *Corrected (TP)*. The two levels will be contrasted with one another to highlight the different behaviour of participants when recognizing and correcting an error (TP) compared to the reference level, i.e., during normal reading of a sentence without an error (TN). Participants were presented with either a source and a test sentence (possible interactions: TP, FP) or a source and a control sentence (possible interactions: TN, FN), never with a source, test, *and* control sentence, i.e., in either case, they saw a source and a (correct (TN) or incorrect (TP)) target sentence. The distinction between test and control is made through the use of *TN* and *TP*. The second predictor we used is the variable *Error Type*, which has the three levels *Misterm, Reg, Func*. To be able to contrast categorical variables in LMERs, we used dummy coding, which compares each level of the two categorical variables *Error Type* and *Corrected* to a reference level. The reference level for the predictor *Error Type* is *Func*. In other words, the two levels of the variable *Error Type,* i.e., *Misterm* and *Reg*, were compared to the reference level *Func* errors. The reference level for the predictor *Corrected* represent, as mentioned before, *True Negatives (TN)*. The two predictors are used to test all dependent variables.

The dependent variables in the LMERs are based on standard eye movement measures (see Table 4). They are Total Reading Times on Source Text Tokens Aligned to Target Text Tokens (TrtS), Total Reading Times on Target Text Token (TrtT), Regression Path Duration on Target Text Token (RPDur), First Pass Duration on Target Text Token (FPDurT) and First Fixation Duration on Target Text Token (FFDur), listed here according to the cognitive processing stage they describe. There are cognitive processes that occur rather automatically, while others are more controlled. "[R]eading requires a coordination between automatic and attention-demanding (control) processing activities" [52], and the same goes

for translation, but in more than one language. This processing dichotomy is also represented in the eye-tracking measures. Early measures represent faster, more automated cognitive processes, while late measures represent slower, more conscious cognitive processes. Total reading time is the sum total of all fixations on a token and is hence considered a measure for slow processes. Regression path durations describe the time that passes from a first fixation on a token, including on prior words to the left of said token, to a fixation on a different token to the right. Regression path duration can represent a range of aspects but is usually viewed as a failure to integrate information which is mediated by re-reading previous text. First pass duration is the sum of all fixation durations starting from the first fixation on a token until the eyes move to a different token. They happen slightly later than first fixation durations, which describe the first contact with a word and are usually seen as an indicator of all the processes leading up to (e.g., visual/phonological, orthographic, or morphological information processing) and including (bilingual) lexical access. Lexical access defines the process of entering the mental lexicon, which can be understood as a database of all the words stored in the mind of an individual (in one or more languages) [53]. If subsequent fixations fall on the same word before a different word is fixated, they are measured with first pass durations. A second, a third or even more fixations may be necessary if the word is long, infrequent and/or difficult to translate.

Table 4. The eye-tracking measures used as dependent variables in the regression models.

Eye-Tracking Measure	Definition
FFDur	first fixation duration is the duration of the first fixation on the token
FPDurT	first pass target token reading duration is the sum of all fixation durations on the target token from the first fixation until the participants looks at a different token
RPDur	regression path duration is the amount of time it took from the first contact with a word until the eyes move on to the right in the text. It includes all regressions to the left.
TrtS	total reading time on source token is the sum of all fixations on the source token(s) aligned to a particular target token for the duration of the entire session
TrtT	total reading time on target token is the sum of all fixations on the target token for the duration of the entire session

The dependent variables were transformed with the natural logarithm to ensure normal distribution. Zeros were excluded for each dependent variable, which means that all tokens that the participants did not look at were excluded from further analysis.

Furthermore, we included the random variables *Participant* and *Token* (i.e., the error token in a test sentence or the correct version of the error token in a control sentence) in all models to account for individual differences that may have had a crucial impact on the data.

Although we had controlled for the length of TT tokens in terms of what had been initially presented to participants, we had no control over modifications they would perform. In other words, TT tokens in the final versions produced by participants might have affected reading times in ways we cannot control, which is why we included word length in characters of the target token (LenT) in the models for TT reading times.

Subsequently, to test for the effect of errors on ST token reading time, we included the number of ST tokens aligned to TT tokens in the models for ST total reading times, because participants were free to edit without restrictions and the target token could therefore not be controlled for. This means that alignment relations between source and target tokens are not exclusively *word to word* but sometimes also *m to n words* relations, which may lead to longer reading times on a ST token if it is aligned to more than one TT token. This will be taken into account by including the number of ST tokens aligned to TT tokens in the models.

For each of the dependent variables, we built three models. We first checked for the main effects of the variable *Corrected*, which were all highly significant, as is logical. A surprising result would

have been if there were no significant differences in eye-movement behaviour between correct tokens left unchanged and wrong tokens that were corrected. This would have made any further analysis pointless. We then checked for main effects of the variable *Error Type* to see whether the error category has an effect on eye movement behaviour, irrespective of whether an error was present or not. Lastly, we looked at interaction effects of the two variables *Corrected* and *Error Type*. The main effects of the LMERs with the predictors *Corrected* and *Error Type* on each dependent variable are presented in Tables 5 and 6:

Table 5. T-values and significance intervals (*p*) for the main effects of *Corrected* on the dependent variables.

Main Effect		**Predictor in All Models**	
Random variables in all models: (1|Part) + (1|TToken)		*Corrected* (corrected token (TP) compared to correct token (TN))	
		t	*p*
Dependent variables in all models:	log (TrtS)	4.74	2.40×10^{-6} ***
	log (TrtT)	5.59	2.93×10^{-8} ***
	log (RPDur)	3.83	$<2 \times 10^{-16}$ ***
	log (FPDurS)	2.09	0.0367 *
	log (FPDurT)	4.32	1.74×10^{-5} ***
	log (FFDur)	4.02	6.26×10^{-5} ***
Additionally:	TrtS		TrtT
	In TrtS models, we included ST tokens aligned to TT tokens, which had a significant effect.		In TrtT models, we included word length in characters of the target token (LenT), which had a significant effect.

*** $p \leq 0.001$; * $p \leq 0.05$.

Table 6. T-values and significance intervals (*p*) for the main effects of *Error Type* on the dependent variables.

Main Effect		**Predictor in All Models**			
Random variables in all models: (1|Part) + (1|TToken)		*Error Type* Reference level: Func			
		Misterm		*Reg*	
		t		*p*	
Dependent variables in all models:	log (TrtS)	5.39	1.27×10^{-7} ***	2.11	0.0358 *
	log (TrtT)	2.08	0.04 *	-	Not sig.
	log (RPDur)	-	Not sig.	1.81	0.0741.
	log (FPDurS)	-	Not sig.	-	Not sig.
	log (FPDurT)	-	Not sig.	-	Not sig.
	log (FFDur)	-	Not sig.	-	Not sig.
Additionally:		TrtS		TrtT	
	In TrtS models, we included ST tokens aligned to TT tokens, which had a significant effect.			In TrtT models, we included word length in characters of the target token (LenT), which had a significant effect.	

*** $p \leq 0.001$; * $p \leq 0.05$.

Since the predictors had more than one level to be contrasted, and since we applied dummy coding (comparing all levels individually to one reference level), additional post-hoc analyses were carried out to assess the statistical significance of pairwise comparisons of each level combination. For this purpose, we used the package *emmeans* [54] to calculate Estimated Marginal Means [55], i.e., to assess the statistical significance of the different marginal means for each pairwise comparison.

4. Results

4.1. Errors Corrected vs. Not Corrected

Since two participants did not complete all 11 session recordings, five texts are missing, resulting in a total of 2397 presented tokens. Table 7 provides an overview of the token distribution in our study:

Table 7. Error counts and *Error Type* distribution of presented tokens: raw counts and percentages.

	Err	Corr	Total
Reg	606 (25.28%)	604 (25.20%)	1210 (50.48%)
Misterm	311 (12.97%)	311 (12.97%)	622 (25.95%)
Func	283 (11.81%)	282 (11.76%)	565 (23.57%)
Total	1200 (50.06%)	1197 (49.94%)	2397 (100%)

Out of 2397 tokens, 1200 were error tokens (one per test sentence) and 1197 were correct tokens (one per control sentence). More precisely, we presented 1210 *Reg* sentences (606 Corr/604 Err), 622 *Misterm* sentences (311 Corr/311 Err) and 565 *Func* sentences (282 Corr/283 Err). We included a higher number of *Reg* sentences to have more statistical power for the LMERs, because we hypothesized that this *Error Type* would be overlooked more often than the other categories, since stylistic changes largely depend on personal preferences. In other words, we anticipated that participants would correct fewer *Reg* errors and that there would be far less overlap between participants regarding which *Reg* errors they would correct. Since we excluded all sentences with an error which were not corrected and since a critical mass of items is necessary in terms of statistical power, we included more sentences with *Reg* errors so that we would end up with sufficient items in this *Error Type* category. Table 8 shows the results of the four possible interactions of the participants with the tokens.

Table 8. True positives, false positives, true negatives, and false negatives of presented tokens: raw counts and percentages.

	Misterm	***Reg***	***Func***	**Total**
TP	230 (73.95%)	109 (17.99%)	123 (43.46%)	462 (38.50%)
FP	81 (26.05%)	497 (82.01%)	160 (56.54%)	738 (61.50%)
TN	273 (87.78%)	493 (81.62%)	235 (83.33%)	1001 (83.66%)
FN	38 (12.22%)	111 (18.38%)	47 (16.67%)	196 (16.37%)

Just over a third (38.5%) of all error tokens were corrected, while 61.5% were left uncorrected. In total, 83.66% of all correct tokens remained untouched, while 16.37% were modified unnecessarily. Participants seem to correct stylistic errors the least (17.99%), which confirms our hypothesis that they are overlooked more often. Almost half (43.5%) of the incorrect function words and nearly three-quarters (73.95%) of mistranslation errors were corrected. For all three categories, more than 80% of the control tokens that should not have been changed remained untouched. It would be interesting to analyse coherent texts instead of single sentences to see to what extent this outcome changes.

4.2. Total Reading Times of Source Text and Target Text Tokens

The first LMERs were fitted to look at the total reading times of ST and TT tokens. As described before, the data were transformed with the natural logarithm to achieve a more or less normal distribution, and then visualized in density plots, i.e., graphical representations of the data distribution estimations of numeric variables. The following density plots show the estimated distribution of total reading times on source and target text tokens (see Figures 5 and 6). The peak in the plot is where most data values are concentrated.

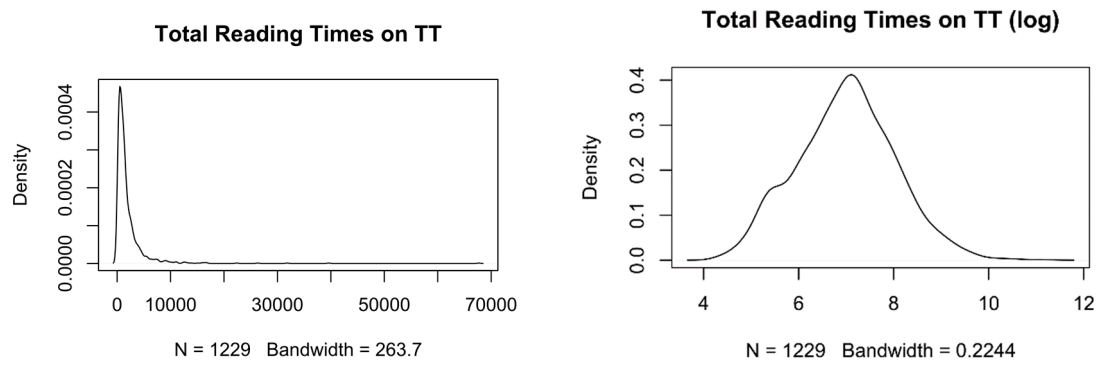

Figure 6. Density (**left**) and log-transformed density (**right**) of total reading times on target text.

We first tested for the effect of *Corrected* on total reading times. What we compare with this variable is, as mentioned above, the eye movement behaviour on error words in the test sentences with that on the 'same' words in the control sentences which are correct. The main effect of *Corrected*, where differences between error tokens and correct tokens are investigated, on TrtS ($\beta = 2.56 \times 10^{-1}$, $SE = 5.40 \times 10^{-2}$, $t = 4.741$, $p < 0.001$) (see Figure 7 left) and on TrtT ($\beta = 3.55 \times 10^{-1}$, $SE = 6.37 \times 10^{-2}$, $t = 5.59$, $p < 0.001$) (see Figure 7 right) was highly significant, such that TrtT and TrtS on error tokens which were corrected (TP) were significantly longer than for correct tokens (TN), which confirms that errors were recognised.

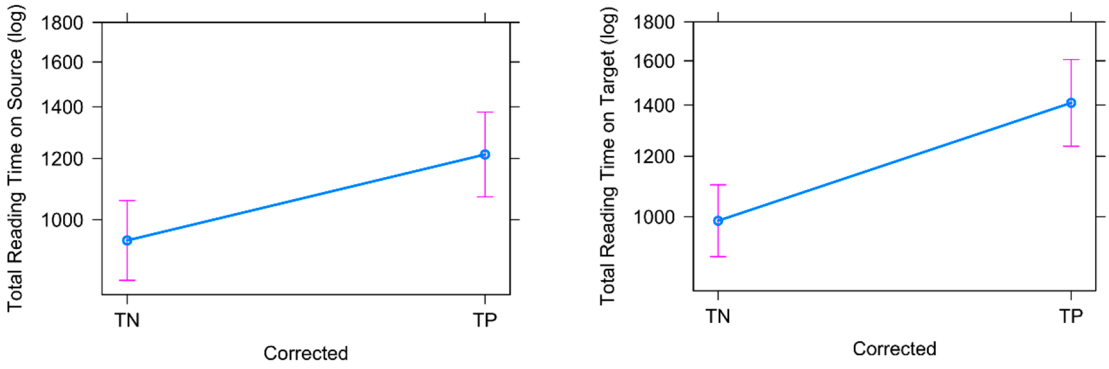

Figure 7. Effect of *Corrected* on total reading times in ST (**left**) and TT (**right**).

Regarding the effect of *Error Types*, we tested to what extent TrtS (see Figure 8 left) differed according to the *Error Type*—irrespective of whether the token was erroneous or correct (both TP and TN tokens were thus included). We found that TrtS (see Figure 8 left) was shortest for the *Error Type Func*, which is not very surprising, since function words are often rather short words. ST tokens aligned to register tokens were read significantly longer than ST tokens aligned to *Func* tokens ($\beta = 0.20$, $SE = 0.10$, $t = 2.11$, $p < 0.05$) and ST tokens aligned to *Misterm* tokens were read significantly longer than *Func* tokens ($\beta = 0.50$, $SE = 0.09$, $t = 5.39$, $p < 0.001$) (see Figure 9 left). Post-hoc tests showed that TrtS was significantly longer for *Misterm* than *Reg* items ($\beta = -0.42$, $SE = 0.09$, $t = -4.64$, $p < 0.0001$).

TrtT (see Figure 8 right) on *Misterm* tokens differed significantly from TrtT on *Func* tokens ($\beta = 0.20$, $SE = 0.10$, $t = 2.08$, $p < 0.05$), TrtT on *Reg* tokens did not. Post-hoc tests showed that TrtT was not significantly different for *Misterm* and *Reg*.

To sum up, TrtS were longer for *Misterm* tokens than for *Reg* tokens and this was most likely driven by erroneous tokens, given that mistranslations/terminology errors require comparison with corresponding ST tokens more often than register or function word errors, which can usually be fixed by simply reading the TT. Register errors might be more difficult to detect in the first place but the source text is not necessarily needed for correction because they only affect the fluency of texts,

i.e., if a register error is corrected, the total reading time in the ST is shorter than for *Misterm* errors. The same goes for function errors, which are listed under fluency in the MQM framework.

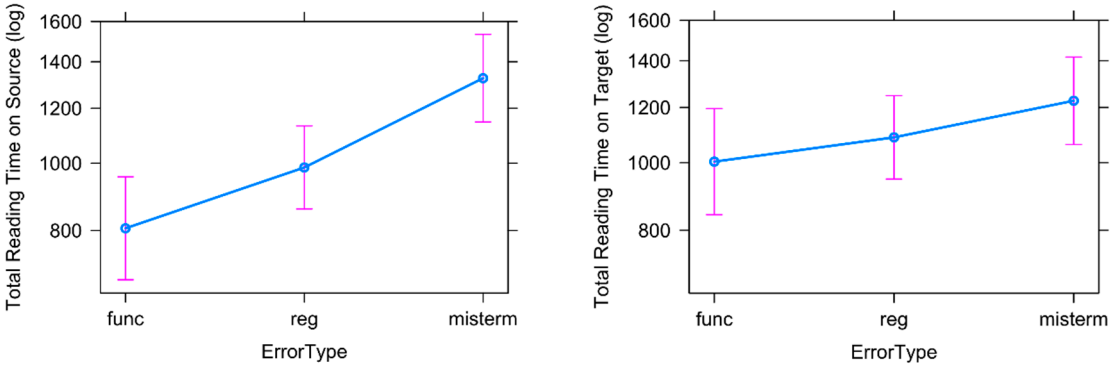

Figure 8. Effect of *Error Type* on total reading times in ST (**left**) and TT (**right**).

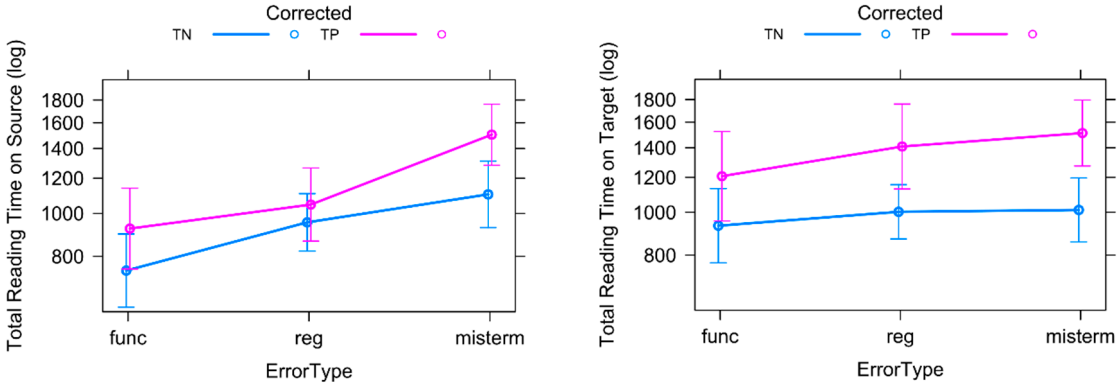

Figure 9. Effect of the interaction *Corrected* with *Error Type* on total reading times in ST (**left**) and TT (**right**).

Next, we tested for interactions between corrected errors and *Error Type* for TrtT and TrtS (aligned to TT tokens). As mentioned above, we included the number of aligned ST tokens as a predictor in all models with total reading time on the source as a dependent variable and word length in characters of target words for models with TrtT as a dependent variable.

The interaction between *Corrected* and *Error Type* on TrtS (see Figure 9 left) was not significant. The post-hoc analysis revealed that ST tokens aligned to TT *Func* tokens were read significantly longer when an error was present than when not ($\beta = -0.22$, $SE = 0.11$, $t = -2.03$, $p < 0.05$) and the same was true for *Misterm* tokens ($\beta = -0.31$, $SE = 0.08$, $t = -3.75$, $p < 0.001$). However, ST total reading times did not differ significantly for *Reg* tokens. Additional post-hoc tests showed that while True Positive (TP) total reading times on the ST for *Reg* errors did not differ significantly from those for *Func* TPs, the TrtS for TP *Misterm* errors was significantly longer than those for TP *Reg* errors ($\beta = -0.36$, $SE = 0.11$, $t = -3.26$, $p < 0.01$).

The interaction between *Corrected* and *Error Type* on TrtT (see Figure 9 right) was not significant either, but the post-hoc analysis revealed that *Func* tokens were read significantly longer if there was an error present than when not ($\beta = -0.26$, $SE = 0.13$, $t = -2.0$, $p < 0.05$), the difference in TrtT for erroneous and correct *Misterm* tokens was highly significant ($\beta = -0.40$, $SE = 0.10$, $t = -4.07$, $p = 0.001$). The effect of a register error on TrtT was also significant ($\beta = -0.34$, $SE = 0.11$, $t = -2.9$, $p < 0.05$). Additional post-hoc analysis showed that the TrtT of correct tokens (TN) did not differ significantly between any of the three *Error Types*. The same was true of corrected tokens (TP).

In sum, professional translators of the DGT required similar amounts of time for the correction of *Func*, *Misterm* and *Reg* errors. Not surprising is the fact that *Misterm* errors led to longer ST reading

times, reflecting the need to compare the TT with the ST. Neither is it surprising that *Reg* errors did not lead to significantly longer ST total reading times—the very nature of register errors, i.e., the fact that the erroneous and correct items were mostly synonymous, meant that a comparison with the source was unnecessary in many cases. However, the fact that the time spent reading the TT for the three *Error Types*, if they were corrected, was the same, and the fact that the interaction between *Error Type* and the *Corrected* variable was not significant for TrtT is surprising. These results suggest that participants either found the correction of all three *Error Types* equally difficult or that they were equally conscientious, double and triple checking every error to the same extent.

4.3. Regression Path Duration on Target Text

In the next step, we fitted LMERs to investigate the regression path duration in the TT. Again, the data were transformed with the natural logarithm to achieve a more or less normal distribution (see Figure 10):

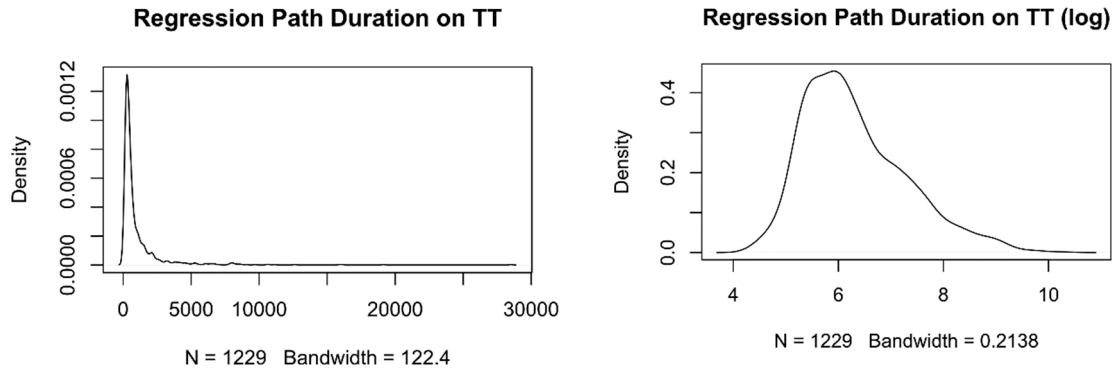

Figure 10. Density plot (**left**) and log-transformed density plot (**right**) for regression path duration on TT.

We found a highly significant main effect of *Corrected* on RPDur in the TT ($\beta = 0.24$, $SE = 0.06$, $t = 3.83$, $p < 0.001$; see Figure 11 left), which means that participants re-read previous TT tokens if they encountered an error.

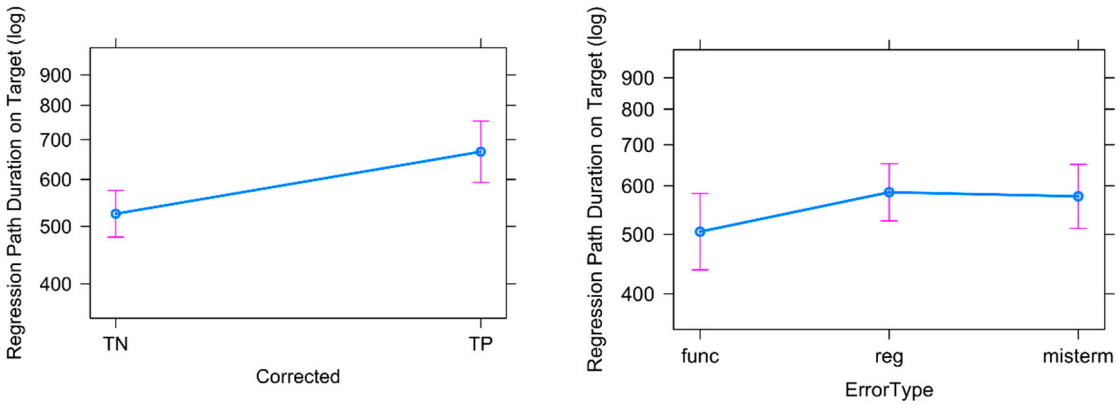

Figure 11. Effect of corrected (**left**) and error type (**right**) on regression path duration on target text.

The main effect of *Error Type* on RPDur in the TT (see Figure 11 right) was marginally significant only for register tokens ($\beta = 0.15$, $SE = 0.08$, $t = 1.81$, $p < 0.08$). The post-hoc analysis did not reveal any significant differences in regression path durations between any of the categories. The effect of the interaction between *Corrected* and *Error Type* on RPDur in the TT (see Figure 12) was not significant. However, the post-hoc analysis showed that erroneous tokens led to longer regression path durations compared to the correct tokens for all *Error Types* (*Func*: $\beta = -0.31$, $SE = 0.12$, $t = -2.74$, $p < 0.01$;

Reg: $\beta = -0.29$, $SE = 0.13$, $t = -2.23$, $p < 0.05$; *Misterm*: $\beta = -0.21$, $SE = 0.10$, $t = -2.12$, $p < 0.05$). Additional contrasts in the post-hoc analysis showed that the RPDur for TP and TN tokens did not differ significantly between different *Error Types*.

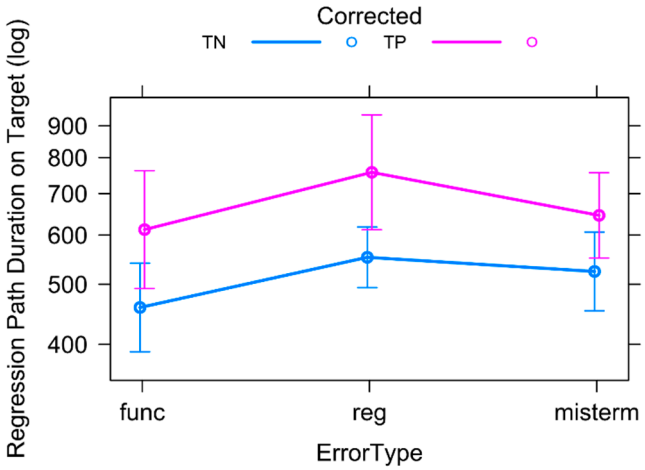

Figure 12. Interaction between *Corrected* and *Error Type* on target text regression path duration.

In conclusion, if professional translators at the DGT encounter an error, they seem to re-read the previous context irrespective of the *Error Type* similarly. In other words, previous context seems to be equally important when recognising *Func, Reg* or *Misterm* errors. This is surprising, given that *Func* errors might be argued to not require previous context as much as *Reg* or *Misterm* errors. What these results suggest is that, in order to correct errors, participants needed the prior context in equal measure across all error categories.

4.4. First Pass Duration (TT)

The next LMERs were fitted to look at first pass durations in the TT, with the data also being log-transformed (see Figure 13).

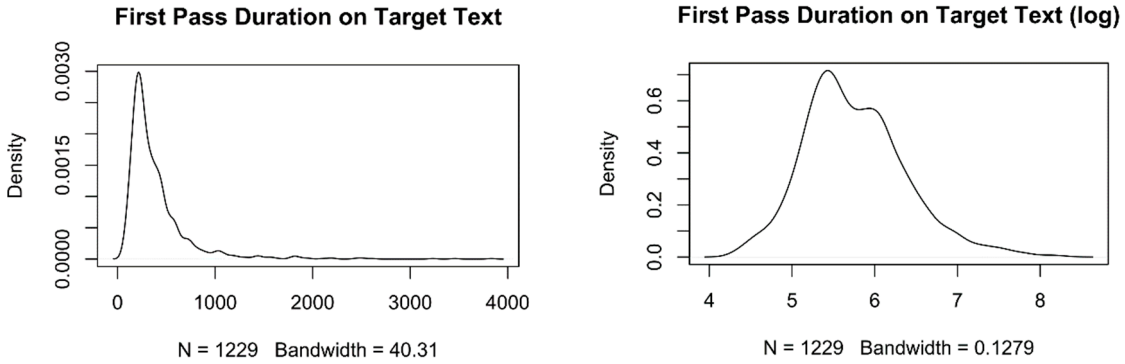

Figure 13. Density plot (**left**) and log-transformed density plot (**right**) for source text first pass duration.

The main effect of *Corrected* (see Figure 14 left) was highly significant ($\beta = 0.17$, $SE = 0.04$, $t = 4.31$, $p < 0.001$). The effect of *Error Type* on FPDurT was not significant and the post-hoc analysis did not reveal any further significant effects.

The interaction between *Corrected* and *Error Type* on FPDurT (see Figure 14 right) was not significant either. The post-hoc analysis showed that first pass durations for *Func* ($\beta = -0.21$, $SE = 0.08$, $t = -2.67$, $p< 0.01$), *Reg* ($\beta = -0.16$, $SE = 0.07$, $t = -2.27$, $p < 0.05$) and *Misterm* ($\beta = -0.15$, $SE = 0.06$, $t = -2.38$, $p < 0.05$) errors were significantly longer when there was an error present than when not. Further post-hoc contrasts showed that only the FPDurT for *Misterm* tokens which contained an error (TP)

was marginally shorter than the FPDurT for *Func* tokens which contained an error (TP) ($\beta = -0.18$, $SE = 0.09$, $t = -2.16$, $p < 0.08$). There was no significant difference in FPDurT between the different *Error Types* for correct tokens (TNs).

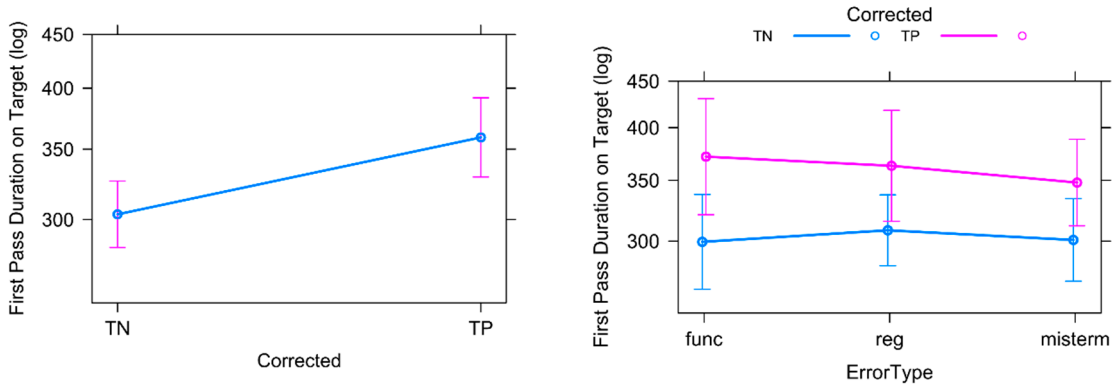

Figure 14. Effect of *Corrected* (**left**) and the interaction *Corrected/Error Type* (**right**) on target text first pass duration.

These results are insofar surprising, as we expected different results for different error categories. In the existing study by [22], there were no early effects for accuracy errors such as sense errors, while here all *Error Types* had an effect on First Pass Reading Times. It would be interesting to investigate whether these results change if we further divide the error categories into errors that require consulting the source text and errors that can be spotted in the target text alone. The fact that *Func* tokens which contained an error (TPs) were read marginally longer than *Misterm* tokens which contained an error (TPs) is the only similarity between the two studies, suggesting that in the early stages, participants lingered slightly longer on the more obvious grammatical incorrectness of *Func* items than the more subtle and semantic differences between erroneous and correct *Misterm* items.

4.5. First Fixation Duration on Target Text

The first fixation durations (FFDur) were log-transformed again (see Figure 15).

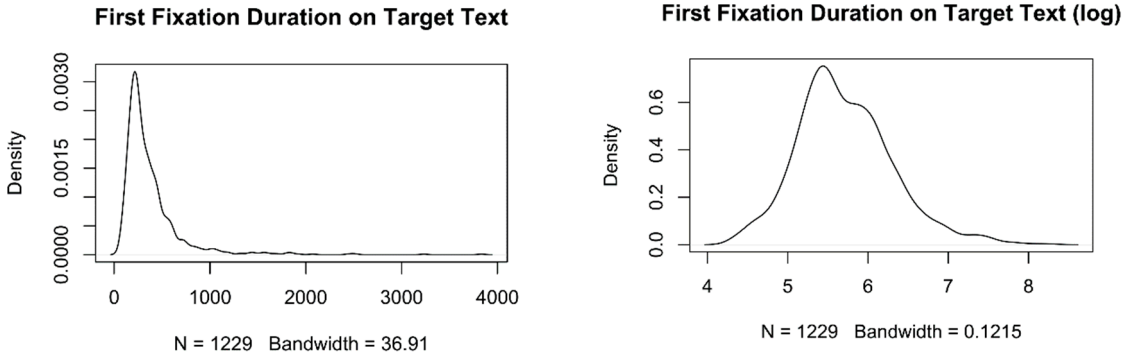

Figure 15. Density plot (**left**) and log-transformed density plot (**right**) for target text first fixation durations.

Looking at first fixation durations on target text tokens, we found a highly significant main effect of *Corrected* on FFDur ($\beta = 1.51 \times 10^{-1}$, $SE = 3.78 \times 10^{-2}$, $t = 4.01$, $p < 0.001$; see Figure 16 left), i.e., corrected error tokens had highly significantly longer first fixation durations than correct tokens.

The main effect of *Error Type* on FFDur (see Figure 16 right) was not significant. The post-hoc analysis did not reveal any further significant effects.

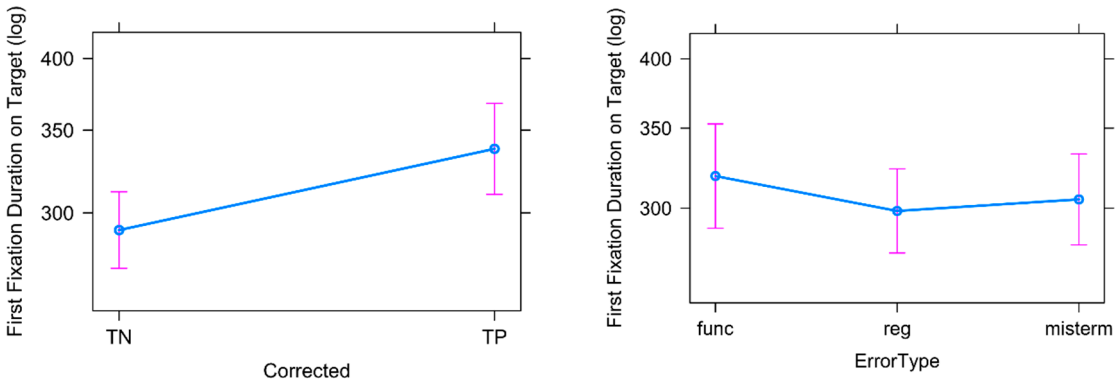

Figure 16. Effect of *Corrected* (**left**) and *Error Type* (**right**) on target text first fixation durations.

The interaction between *Corrected* and *Error Type* on FFDur (see Figure 17) was not significant. The post-hoc analysis revealed that first fixation durations for *Func* ($\beta = -0.21$, $SE = 0.08$, $t = -2.70$, $p < 0.007$), *Reg* ($\beta = -0.18$, $SE = 0.07$, $t = -2.55$, $p < 0.05$), and *Misterm* errors ($\beta = -0.12$, $SE = 0.06$, $t = -2.08$, $p < 0.05$) were significantly longer when there was an error present than when not. Additional post-hoc analyses did not reveal any significant differences between different *Error Types*.

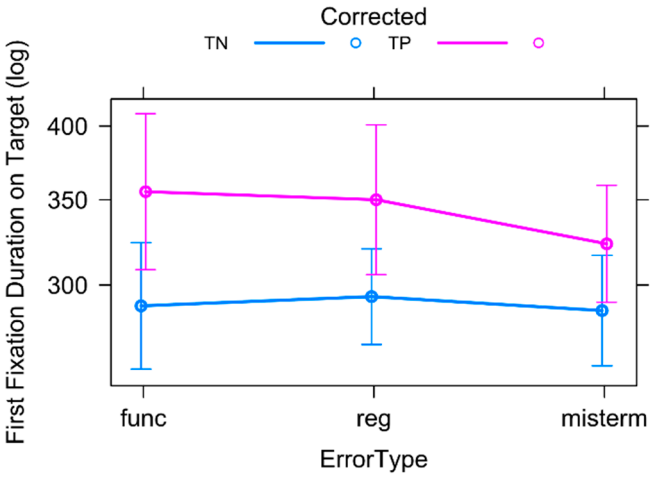

Figure 17. Interaction effect of *Corrected* and *Error Type* on first fixation duration on target text.

The fact that all three error categories had an effect on first fixation durations is very surprising. First fixation durations typically represent word recognition processes, which "involves retrieving information about a word's [. . .] meaning from its printed form" [56] up to and including lexical access, i.e., accessing the mental lexicon to retrieve information about words, and the fact that errors had an effect on this early measure suggests that these participants are highly tuned to quickly fixing all linguistic or cultural conventions that these errors had contravened and that, even for aspects of stylistic errors, for this group of participants, many of the processes regarding the recognition of errors occur to a large extent automatically and with little conscious control over the mechanisms which lead to error recognition.

4.6. Eye–Key Span

The eye–key span [57] describes the temporal distance between the first contact with a word and the first keystroke which contributes to the typing of the translation of this word. This kind of measure typically represents very slow processes: many intervening processes take place between a first contact with a word and the typing of its translation—from re-reading immediate context, to possible typing at different places in the text and re-reading of more distant text material. A shorter eye–key span can be

interpreted, on the most general level, as requiring less cognitive effort and more specifically as being an indicator of the certainty with which the typing is carried out—the more temporal immediacy of the typing in relation to a first contact, the easier this translation is and the more certain the translator is. The authors of [58] argue that translators activate translations very early during the translation process and that these early automatic processes serve as the basis for later text production processes. We tested this hypothesis by analysing the effect of the first fixation duration on the eye–key span and the interaction between *Error Type* and first fixation duration. The rationale here was that the result of the more or less automatic error recognition processes during a first contact with the word should interact with later processes and this interaction might be modulated by *Error Type* regarding the participants certainty. Specifically, large effects during the early processes should lead to shorter eye–key spans than small effects in the early stages of error recognition.

For this last analysis, we only used the corrected error tokens (TP), because there are of course no keystrokes associated with correct tokens which were not modified. The eye–key span was not normally distributed, which is why we log-transformed the data (see Figure 18). The main effect of FFDur on the eye–key span was significant ($\beta = -0.51$, $SE = 0.16$, $t = 3.23$, $p < 0.01$). The main effect of *Error Type* was not, and neither was the interaction between *Error Type* and FFDur. However, the post-hoc analysis revealed that the effect of FFDur on the eye–key span was significant for corrected *Misterm* tokens ($\beta = -0.60$, $SE = 0.21$, $t = 2.84$, $p < 0.01$) and for corrected *Func* tokens ($\beta = -0.63$, $SE = 0.32$, $t = 1.99$, $p < 0.05$). It was not significant for corrected *Reg* tokens (see Figure 19).

Although the interaction between first fixation durations and error types for the eye–key span was not significant, the post-hoc analysis revealed the only dissociation between *Error Type* categories for corrected items in our data. The dissociation was such that if an error was spotted early and if this effect was strong, then its correction occurred more immediately than for those errors which were recognised early but did not lead to large early effects. This was only the case for *Misterm* and *Func* errors. For *Reg* errors, there was no relation between the early and the very late effects in the eye–key span. This pattern of results might suggest that the severity of the error and/or the certainty regarding the correction of *Misterm* and *Func* errors which are recognised during the early processes has an effect on later processes: large early effects have an effect on the later measure (eye–key span). The previous analyses for all other eye movement measures have shown that early effects typically, and particularly if they are strong, show up in the later measures (slower processes). This suggests that the early, more automatic processes feed into the slower, later processes. The eye–key span is a much later and slower process than that represented by total reading times, because potentially, many more different processes intervene. These results hint at the importance of the early, more automatic processes, and, in this regard, differences between the behaviour of professionals on the one hand and students on the other hand promise to reveal important effects.

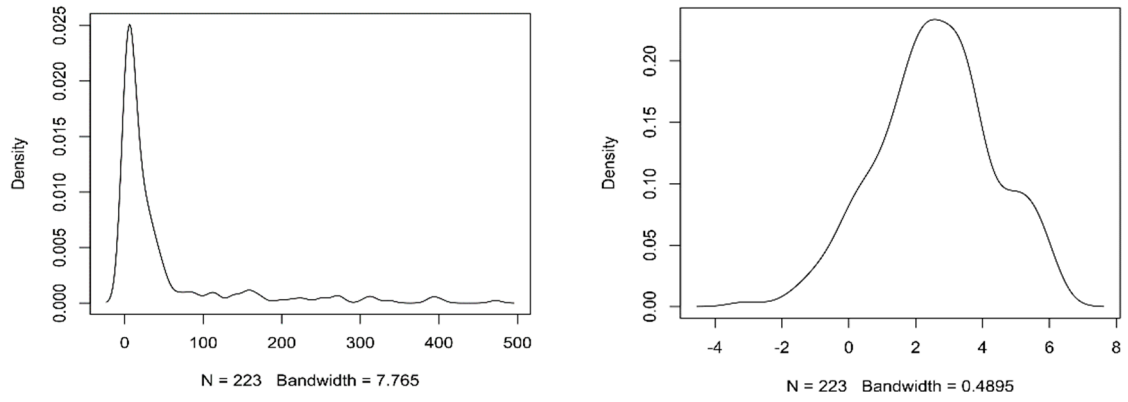

Figure 18. Density plot (**left**) and log-transformed density plot (**right**) for eye–key span.

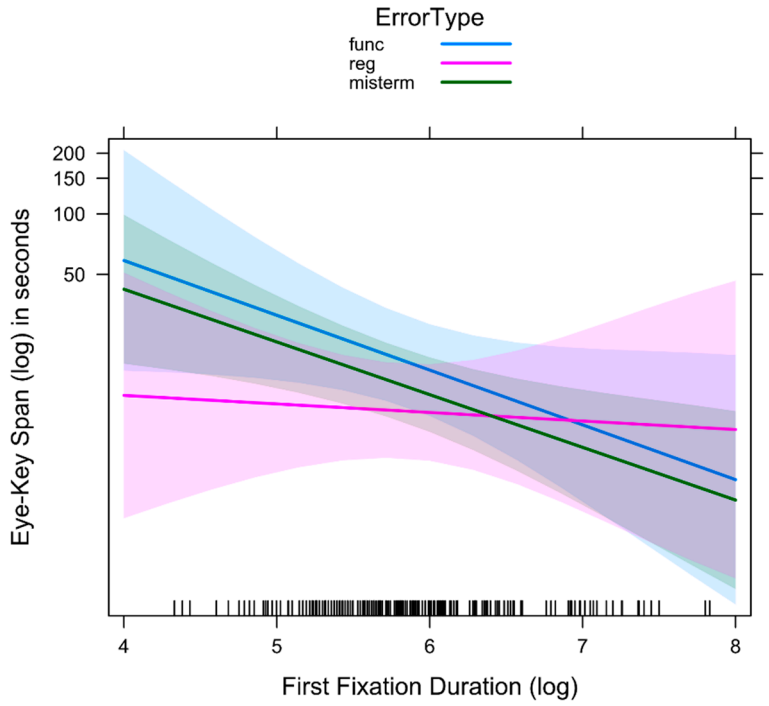

Figure 19. Interaction between log-transformed first fixation durations and *Error Type* for eye–key span (the time between first visual contact with a target word and the first keystroke that contributes to the correction of this token).

5. Conclusions

This paper reports on two consecutive studies: (1) a published corpus analysis of PE changes and their revisions performed by DGT professionals to explore the quality of NMT outputs and NMTPE of authentic texts by conducting automatic error annotations with Hjerson, as well as a subsequent manual annotation using the MQM framework to clean up the automatic annotation and to gain a deeper understanding of its subclasses [35], and (2), based on the results of this corpus analysis, a controlled eye-tracking experiment to analyse how professional translators recognize and correct errors.

Previous studies on NMT quality reported omissions and mistranslations/incorrect lexis as problematic NMT error categories, e.g., [13,17–19], and a similar picture is painted in our analysis of the DGT corpus, where lexical errors, and particularly stylistic/register errors, function words, mistranslations, and terminology errors are the most common error types—both in the machine translation output and the post-edits, which might be explained by the priming effect of machine translations. The most prominent error categories, i.e., mistranslations, terminology errors, function words and stylistic/register errors, were further analysed in a key-logging and eye-tracking experiment to gain more insights regarding their effect on PE effort indicators such as the eye-tracking measures.

Similar work has been performed before. Using eye-tracking and key-logging as well, [10] investigated how different MT error types impact the PE effort and found out that "most post-editing effort indicators (product as well as process) are influenced by machine translation quality, but that different error types affect different post-editing effort indicators" [10]. However, because they worked on a sentence level and their sentences often contained more than one error of various categories, which makes it difficult to isolate single processes, they cannot tie their results to specific error words that cause an increase in PE effort. Since they did not manipulate their texts, they can only study the effect of the error weights and not the effect of errors compared to the effect of correct versions. Furthermore, we worked with authentic DGT texts containing authentic errors that had been created and corrected in real-life scenarios by highly multilingual and proficient DGT translation experts with a comparable profile, who also took part in our eye-tracking experiment.

We expected different eye movement behaviour when detecting and correcting different error categories, but the DGT professionals recognized all error categories similarly early and they do not seem to prioritize certain error categories over others. There is almost no difference between how they treat mistranslations that require consultation of the source text and how they treat, e.g., stylistic changes, where the target text provides sufficient information. Only in the relationship between the earliest and a very late measure was there a hint that these participants dissociate between functional word errors and mistranslations on the one hand and stylistic errors on the other. The severity of errors and associated certainty regarding the correction of mistranslations and functional word errors as recognised early during the process is fed into the later processes slightly better than for stylistic errors. However, the participants studied here detected mistranslations, terminology errors as well as stylistic errors during early processes, represented by measures such as first fixation duration and first pass duration, which indicates that the participants in this study recognized all error categories automatically to a large extent.

Furthermore, DGT translators spend more time in the source text if needed, e.g., for mistranslation and terminology errors, while for stylistic errors, participants spent the same amount of time in the source text irrespective of whether there was a mistake or not, which indicates that they prioritize some reading strategies over others depending on what is needed in a specific situation.

To the best of our knowledge, this is the first study to explore the effects of NMT error types on eye movements during professional post-editing based on the results of a preceding corpus analysis. Future research directions involve a comparison of professional behaviour with that of a less experienced group, i.e., translation students from our faculty. Both [10] and [22] found differences between professionals and students. We also expect to find significant effects of the variable *Status*, i.e., translation student or DGT professional, which will allow us to model the translation expertise more precisely than with the data obtained in this study alone. As mentioned in the introduction, the mechanisms which develop with exposure to the task are the equivalent of what QE approaches attempt to model. We try to model these mechanisms in humans by comparing the behaviour of two groups with different translation expertise. We have already recorded the data of 30 student participants the exact same way we recorded the data of the 30 DGT professionals. The only difference is that the sessions took place at our faculty instead of at the DGT. Again, the comparison of professionals and students is crucial, since behavioural differences between the two groups will highlight the effect of professional expertise, i.e., the mechanisms that develop over time with exposure to the task, more clearly.

Supplementary Materials: The supplementary materials are available online at http://www.mdpi.com/2227-9709/6/3/41/s1.

Author Contributions: Conceptualization, S.H.-S. and M.S.; methodology, S.H.-S. and M.S.; validation, S.H.-S. and M.S.; formal analysis, J.V. and M.S.; investigation, J.V. and M.S.; resources, M.S.; data curation, J.V. and M.S.; writing—original draft preparation, J.V. and M.S.; writing—review and editing, S.H.-S. and M.S.; visualization, J.V. and M.S.; supervision, S.H.-S. and M.S.; project administration, S.H.-S. and M.S.

Funding: This research received no external funding.

Acknowledgments: We thank the German Department of the DGT for providing us with data, the professional translators who took part in our experiment, and the reviewers and editors of this special issue for their constructive remarks.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Chartier-Brun, P.; Mahler, K. Machine Translation and Neural Networks for a Multilingual EU. Available online: http://zerl.uni-koeln.de/chartierbrun-mahler-2018-machine-translation-eu.html (accessed on 30 April 2019).
2. 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., Ed.; Springer: New York, NY, USA, 2018; pp. 39–68.

3. ISO 17100. *Translation Services—Requirements for Translation Services*; Technical Committee ISO/TC37: Geneva, Switzerland, 2015.
4. DIN ISO 18587. *Translation Services—Post-Editing of Machine Translation Output—Requirements*; Technical Committee ISO/TC 37/SC 5; Beuth: Berlin, Germany, 2017.
5. Biel, Ł. Quality in Institutional Eu Translation: Parameters, Policies and Practices. *Qual. Asp. Inst. Transl.* **2017**, *8*, 31.
6. Svoboda, T. Translation Manuals and Style Guides as Quality Assurance Indicators: The Case of The European Commission'S Directorate-General for Translation. *Qual. Asp. Inst. Transl.* **2017**, *8*, 75.
7. Wagner, E. Quality of Written Communication in a Multilingual Organisation. *Termin. Et Trad.* **2000**, *1*, 16.
8. Xanthaki, H. The Problem of Quality in EU Legislation: What on Earth is Really Wrong? *Common. Mark. Law Rev.* **2001**, *38*, 651–676. [CrossRef]
9. Daems, J.; Vandepitte, S.; Hartsuiker, R.; Macken, L. The Impact of Machine Translation Error Types on Post-Editing Effort Indicators. In Proceedings of the Fourth Workshop on Post-Editing Technology and Practice, Miami, FL, USA, 30 October–3 November 2015; p. 15.
10. Daems, J.; Vandepitte, S.; Hartsuiker, R.J.; Macken, L. Identifying the Machine Translation Error Types with the Greatest Impact on Post-editing Effort. *Front. Psychol.* **2017**, *8*, 1282. [CrossRef] [PubMed]
11. Daems, J.; Macken, L.; Vandepitte, S. Quality as the sum of its parts: A two-step approach for the identification of translation problems and translation quality assessment for HT and MT+PE. In Proceedings of the MT Summit XIV Workshop on Post-editing Technology and Practice, Nice, France, 2–6 September 2013; pp. 63–71.
12. Lommel, A. Multidimensional Quality Metrics Definition. 2014. Available online: http://www.qt21.eu/mqm-definition/definition-2015-06-16.html. (accessed on 15 March 2019).
13. Bentivogli, L.; Bisazza, A.; Cettolo, M.; Federico, M. Neural versus phrase-based MT quality: An in-depth analysis on English–German and English–French. *Comput. Speech Lang.* **2018**, *49*, 52–70. [CrossRef]
14. Bentivogli, L.; Bisazza, A.; Cettolo, M.; Federico, M. Neural versus Phrase-Based Machine Translation Quality: A Case Study. *arXiv* **2016**, arXiv:1608.04631.
15. Daems, J. *A Translation Robot for Each Translator? A Comparative Study of Manual Translation and Post-Editing of Machine Translations: Process, Quality and Translator Attitude*; Ghent University: Ghent, Belgium, 2016.
16. Lacruz, S. Pauses and Cognitive Effort in Post-Editing. In *Post-editing of Machine Translation Processes and Applications*; O'Brien, S., Winther Balling, L., Carl, M., Simard, M., Specia, L., Eds.; Cambridge Scholars Publishing: Newcastle upon Tyne, UK, 2014; pp. 246–272.
17. Castilho, S.; Moorkens, J.; Gaspari, F. *A Comparative Quality Evaluation of PBSMT and NMT Using Professional Translators*; Cambridge Scholars Publishing: Newcastle upon Tyne, UK, 2017; p. 18.
18. Klubička, F.; Toral, A.; Sánchez-Cartagena, V.M. Quantitative Fine-Grained Human Evaluation of Machine Translation Systems: A Case Study on English to Croatian. *Mach. Transl.* **2018**, *32*, 195–215. [CrossRef]
19. Toral, A.; Sánchez-Cartagena, V.M. A Multifaceted Evaluation of Neural versus Phrase-Based Machine Translation for 9 Language Directions. *arXiv* **2017**, arXiv:1701.02901.
20. Moorkens, J.; Walker, C.; Federici, F.M. Eye Tracking as a measure of cognitive effort for post-editing of machine translation. In *Eye Tracking and Multidisciplinary Studies on Translation*; John Benjamins: Amsterdam, The Netherlands, 2018.
21. Koponen, M.; Salmi, L.; Nikulin, M. A product and process analysis of post-editor corrections on neural, statistical and rule-based machine translation output. *Mach. Transl.* **2019**, *33*, 61–90. [CrossRef]
22. Schaeffer, M.; Nitzke, J.; Tardel, A.; Oster, K.; Gutermuth, S.; Hansen-Schirra, S. Eye-tracking revision processes of translation students and professional translators. *Perspectives* **2019**, 1–15. [CrossRef]
23. Mertin, E. *Prozessorientiertes Qualitätsmanagement im Dienstleistungsbereich Übersetzen*; Lang: Frankfurt am Main, Germany, 2006.
24. Carl, M.; Schaeffer, M.; Bangalore, S. The CRITT translation process research database. In *New Directions in Empirical Translation Process Research—Exploring the CRITT TPR-DB*; Carl, M., Schaeffer, M., Bangalore, S., Eds.; Springer: London, UK, 2016; pp. 13–56.
25. Specia, L.; Shah, K. Machine Translation Quality Estimation: Applications and Future Perspectives. In *Translation Quality Assessment*; Moorkens, J., Castilho, S., Gaspari, F., Doherty, S., Eds.; Springer International Publishing: Cham, Switzerland, 2018; Volume 1, pp. 201–235.

26. Wang, J.; Fan, K.; Li, B.; Zhou, F.; Chen, B.; Shi, Y.; Si, L. Alibaba Submission for WMT18 Quality Estimation Task. In Proceedings of the Third Conference on Machine Translation: Shared Task Papers, Brussels, Belgium, 31 October–1 November 2018; pp. 809–815.
27. Shimanaka, H.; Kajiwara, T.; Komachi, M. RUSE: Regressor Using Sentence Embeddings for Automatic Machine Translation Evaluation. In Proceedings of the Third Conference on Machine Translation: Shared Task Papers, Brussels, Belgium, 31 October–1 November 2018; pp. 751–758.
28. Specia, L. Exploiting Objective Annotations for Measuring Translation Post-editing Effort. In Proceedings of the 15th conference of the European Association for Machine Translation, Leuven, Belgium, 30–31 May 2011; p. 8.
29. Specia, L.; Cancedda, N.; Turchi, M.; Cristianini, N.; Dymetman, M. Estimating the sentence-level quality of machine translation systems. In Proceedings of the 13th Conference of the European Association for Machine Translation, Barcelona, Spain, 14–15 May 2009; p. 8.
30. Luong, N.Q.; Besacier, L.; Lecouteux, B. LIG System for Word Level QE task at WMT14. In Proceedings of the Ninth Workshop on Statistical Machine Translation, Baltimore, MA, USA, 26–27 June 2014; pp. 335–341.
31. Scarton, C.; Specia, L. Document-level translation quality estimation: Exploring discourse and pseudo-references. In Proceedings of the 17th Annual Conference of the European Association for Machine Translation, Dubrovnik, Croatia, 16–18 June 2014; pp. 101–108.
32. Graham, Y.; Ma, Q.; Baldwin, T.; Liu, Q.; Parra, C.; Scarton, C. Improving Evaluation of Document-level Machine Translation Quality Estimation. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Short Papers, Valencia, Spain, 3–7 April 2017; Volume 2, pp. 356–361.
33. Popović, M. Hjerson: An Open Source Tool for Automatic Error Classification of Machine Translation Output. *Prague Bull. Math. Linguist.* **2011**, *96*, 59–67. [CrossRef]
34. Vilar, D.; Xu, J.; D'Haro, L.F.; Ney, H. Error Analysis of Statistical Machine Translation Output. In Proceedings of the presented at the International Conference on Language Resources and Evaluation, Genoa, Italy, 22–28 May 2006; p. 7.
35. Vardaro, J.; Schaeffer, M.; Hansen-Schirra, S. Comparing the Quality of Neural Machine Translation and Professional Post-Editing. In Proceedings of the QoMEX, Berlin, Germany, 5–7 June 2019; pp. 1–3.
36. Popović, M.; Ney, H. Word error rates: Decomposition over Pos classes and applications for error analysis. In Proceedings of the Second Workshop on Statistical Machine Translation—StatMT '07, Prague, Czech Republic, 23 June 2007; pp. 48–55.
37. Tezcan, A.; Hoste, V.; Macken, L. SCATE Taxonomy and Corpus of Machine Translation Errors. In *Trends in E-Tools and Resources for Translators and Interpreters*; Corpas Pastor, G., Durán-Muñoz, I., Eds.; Brill: Leiden, The Netherlands, 2017.
38. Lommel, A. *A New Framework for Translation Quality Assessment*; Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI): Kaiserslautern, Germany, 2014.
39. Weingarten, E.; Chen, Q.; McAdams, M.; Yi, J.; Hepler, J.; Albarracín, D. From primed concepts to action: A meta-analysis of the behavioral effects of incidentally presented words. *Psychol. Bull.* **2016**, *142*, 472–497. [CrossRef] [PubMed]
40. Green, S.; Heer, J.; Manning, C.D. The efficacy of human post-editing for language translation. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems—CHI '13, Paris, France, 27 April–2 May 2013; p. 439.
41. Čulo, O.; Nitzke, J. Patterns of Terminological Variation in Post-editing and of Cognate Use in Machine Translation in Contrast to Human Translation. In Proceedings of the 19th Annual Conference of the European Association for Machine Translation, Riga, Latvia, 30 May–1 June 2016; p. 9.
42. Farrell, M. Machine Translation Markers in Post-Edited Machine Translation Output. In Proceedings of the 40th Conference Translating and the Computer, London, UK, 15–16 November 2018; pp. 50–59.
43. Toral, A. Post-editese: An Exacerbated Translationese. *arXiv* **2019**, arXiv:1907.00900.
44. Tupper, D.E.; Cicerone, K.D. Introduction to the Neuropsychology of Everyday Life. In *The Neuropsychology of Everyday Life: Assessment and Basic Competencies, Foundations of Neuropsychology*; Tupper, D.E., Cicerone, K.D., Eds.; Springer: Boston, MA, USA, 1990; Volume 2.

45. García, A.M.; Ibáñez, A.; Huepe, D.; Houck, A.L.; Michon, M.; Lezama, C.G.; Chadha, S.; Rivera-Rei, Á. Word reading and translation in bilinguals: The impact of formal and informal translation expertise. *Front. Psychol.* **2014**, *5*, 1302. [CrossRef] [PubMed]
46. Carl, M. Translog-II: A Program for Recording User Activity Data for Empirical Translation Process Research. *Int. J. Comput. Linguist. Appl.* **2012**, *3*, 136–153.
47. Germann, U. Yawat: Yet another word alignment tool. In Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies Demo Session—HLT '08, Columbus, OH, USA, 15–20 June 2008; pp. 20–23.
48. R Core Team. *R: A Language for Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2018.
49. Bates, D.; Mächler, M.; Bolker, B.; Walker, S. Fitting Linear Mixed-Effects Models Using lme4. *J. Stat. Softw.* **2015**, *67*, 48. [CrossRef]
50. Kuznetsova, A.; Brockhoff, P.B.; Christensen, R.H.B. lmerTest Package: Tests in Linear Mixed Effects Models. *J. Stat. Softw.* **2017**, *82*, 26. [CrossRef]
51. Fox, J.; Weisberg, S. Visualizing Fit and Lack of Fit in Complex Regression Models with Predictor Effect Plots and Partial Residuals. *J. Stat. Softw.* **2018**, *87*, 1–27. [CrossRef]
52. Walczyk, J.J. The Interplay Between Automatic and Control Processes in Reading. *Read. Res. Q.* **2000**, *35*, 554–566. [CrossRef]
53. Dijkstra, T. Bilingual Visual Wod Recognition and Lexical Access. In *Handbook of Bilingualism: Psycholinguistic Approaches*; Kroll, J.F., De Groot, A.M.B., Eds.; Oxford University Press: Oxford, UK, 2009.
54. Lenth, R. *Emmeans: Estimated Marginal Means, aka Least-Squares Means*; R Package Version 1.3.4; The Comprehensive R Archive Network: Vienna, Austria, 2019.
55. Searle, S.R.; Speed, F.M.; Milliken, G.A. Population Marginal Means in the Linear Model: An Alternative to Least Squares Means. *Am. Stat.* **1980**, *34*, 216–221.
56. Snowling, M.; Hulme, C. Editorial Part I. In *the Science of Reading: A handbook*; Snowling, M., Hulme, C., Eds.; Blackwell Publishing: Hoboken, NJ, USA, 2005.
57. Dragsted, B.; Shreve, G.M.; Angelone, E. Coordination of reading and writing processes in translation: An eye on uncharted territory. *Transl. Cogn.* **2010**, *15*, 41.
58. Schaeffer, M.; Carl, M. Shared representations and the translation process: A recursive model. *Transl. Interpret. Stud.* **2013**, *8*, 169–190.