Predicting the Number of Errors in Human Translation Using Source Text and Translator Characteristics

Haruka Ogawa ogawaha@earlham.edu
Department of Languages and Cultures, Earlham College, Richmond IN, 47374, USA

Abstract

Translation quality and efficiency are of great importance in the language services industry, which is why production duration and error counts are frequently investigated in Translation Process Research. However, a clear picture has not yet emerged as to how these two variables can be optimized or how they relate to one another. In the present study, data from multiple English-Japanese translation sessions is used to predict the number of errors per segment using source text and translator characteristics. An analysis utilizing zero-inflated generalized linear mixed effects models revealed that two source text characteristics (syntactic complexity and the proportion of long words) and three translator characteristics (years of experience, the time translators spent reading a source text before translating, and the time translators spent revising a translation) significantly influenced the number of errors. Furthermore, a lower proportion of long words per source text sentence and more training led to a significantly higher probability of error-free translation. Based on these results, combined with findings from a previous study on production duration, it is concluded that years of experience and the duration of the final revision phase are important factors that have a positive impact on translation efficiency and quality.

1 Context

In the language services industry, prompt delivery of an accurate translation is greatly appreciated. However, time and quality are often a trade-off, which is a substantial concern for many translators (Mossop, 2014). Although neither human nor machine can create a "perfect" translation instantly, it is important to identify which factors lead to speedy production and high quality. Translation Process Research (TPR) can shed light on such an essential aspect of translation.

In TPR, efficiency has often been investigated with respect to source text (ST) difficulty and the different levels of expertise possessed by translators (i.e., what distinguishes professional translators from non-professionals, such as student translators or language learners). For example, Sharmin et al. (2008) revealed that more difficult texts attracted longer gaze time, and Dragsted (2005) found that difficult STs slowed down production time. Interestingly, in Dragsted's study, professionals tended to fall back on more novice-like behavior when they were engaged in difficult STs, while professionals exhibited exceptional performance when STs were easy. Moreover, research has shown that professional translators produce translations faster than student translators (Dragsted, 2005; Jakobsen and Jensen, 2008). Although the differences in translator behavior based on expertise and ST difficulty are not always statistically significant (see Hvelplund, 2011), the findings in TPR in general support that time efficiency is influenced by the nature of the ST and certain translator characteristics.

While efficiency is relatively easy to define, translation quality is not due to its multi-faceted nature. Product quality can be measured in various ways, for which Garvin (1984) formulated different approaches: the transcendent, product-based, user-based, manufacturing-based, and value-based approaches. In addition to the quality inherent in the product itself, how clients perceive the product is crucial in translation. Indeed, some clients prioritize cost and time over quality. Such being the case, it is hard to reach a consensus as to which aspect of translation quality should be prioritized, although this topic is actively debated in translation industry (Fields et al., 2014).

Quality measurement also poses problems in translation research, though scholars have attempted to take industry perspectives into account. For instance, Colina (2009) introduces a functionalist translation assessment tool that focuses on user points of view. Within the CRITT TPR-DB community,[1] the Multidimensional Quality Metrics (MQM) framework (Lommel, 2018) is often utilized. The CRITT TPR-DB makes it possible to annotate errors using a scheme based on MQM, on a platform called YAWAT (see Germann, 2008; Carl et al., 2016). Although quality measurement based on error typologies such as those made available through MQM has some disadvantages (see O'Brien, 2012; Daems et al., 2013), the error-based assessment of translation can be useful when accuracy is seen as vital (Kivilehto and Salmi, 2017). Such an assessment is also extremely beneficial to TPR in that it offers clarity and consistency to the field.

The complexity of investigating translation quality makes it difficult to fully capture the trade-off or interplay between production time and quality, especially in human translation (HT). However, some interesting findings have been reported. For example, Daems et al. (2016) examined the use of external resources during HT and post-editing (PE) and found that the overall production time of HT was significantly higher than PE due to the increased time spent on external resources in HT. They also revealed that the overall quality was influenced by the time spent using external resources and that, in HT, the overall error score was lower when the participants consulted external resources for a longer period of time (Daems et al., 2016). In this specific experiment, it seems that time and quality were in fact a trade-off. However, it is still unknown at this point whether this is the case with HT without external resources and/or in different language pairs.

The present study attempts to further elucidate the relationship between production time and translation quality using English-Japanese translation. The research question is: Can we predict the quality of translation based on characteristics of the ST and of individual translators? Here, the quality of translation is operationalized as number of errors, which has been correlated with several process metrics used as indicators of cognitive effort (Vanroy, 2021). A statistical method called zero-inflated generalized linear mixed models (ZIGLMMs) will be utilized, which nicely handles count data skewed by a large number zeros. By doing so, this study aims to identify which characteristics of a ST or a given translator potentially influence translation quality and efficiency.

In the following, Section 2 contains a description of the data; Section 3, the results of statistical analyses. In Section 4, the overall result will be discussed along with some findings from Ogawa (2021), where production duration was predicted by text and translator characteristics, in order to gain a better understanding of the relationship between ST and translator

[1] CRITT TPR-DB stands for Center for Research and Innovation in Translation and Translation Technology Translation Process Research Database. Behavioral and textual data from translation experiments is publicly available, and a list of publications utilizing this database is accessible at https://sites.google.com/site/centretranslationinnovation/tpr-db-publications?authuser=0.

characteristics and translation quality and efficiency.

2 Data and Methodology

The data used here was originally extracted from the ENJA15 project from the CRITT TPR-DB, in which 39 participants translated two out of six STs from scratch. In the present study, there were approximately 13 different translations for each ST.[2] The ".sg tables" from the CRITT TPR-DB were utilized, where the participants' textual and behavioral (i.e., typing and gaze) data is organized in a way that researchers can analyze it at the segment (i.e., sentence) level.

Errors were manually annotated and counted.[3] In doing so, although the unit of analysis was at the segment level, ST and TT did not necessarily have segment-level equivalence. In fact, some translators did divid one ST segment into two TT segments or combine two ST segments into one TT segment. The number of errors was approximated by the number of content words (i.e., nouns, verbs excluding auxiliary verbs, adjectives, and subordinating conjunctions) in the alignment group on the ST side. For example, a participant translated "was imprisoned" as 逮捕_さ_れ_まし_た (literally "was arrested"). This Japanese translation was morphologically analyzed and divided, as marked by underscores, into five tokens, and yet was aligned to "was imprisoned" as a group. In this case, there was only one content word on the ST side of this alignment group, and therefore, only one error was counted despite multiple TT tokens. This method of counting errors roughly but consistently quantified the severity of errors without judging them subjectively and dichotomously (e.g., minor versus critical).

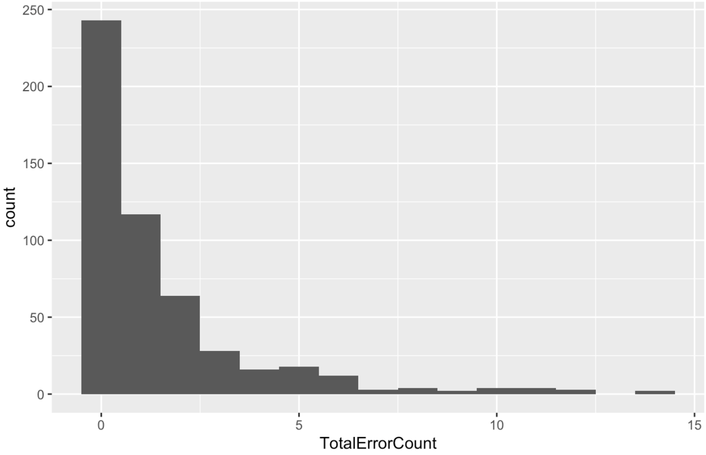


Figure 1: Distribution of TotalErrorCount

The resulting variable, named TotalErrorCount, ranged from 0 to 14 errors. 243 out of 520 segments had zero errors, and 117 segments only had one. As Figure 1 shows, the data was zero-inflated and overdispersed (i.e., the variance is greater than the mean). ZIGLMMs were utilized to handle this skewed data, which have two separate parts. The first part is a count model, which can be interpreted in the same manner as general linear mixed effect models. The count model explains what increases the number of errors. The other part is called a zero-inflated (ZI) model, whose interpretation is equivalent to a logistic regression. It calculates the

[2] See Ogawa (2021) for a more detailed description of the data analyzed.

[3] The errors were originally classified into four categories (i.e., mistranslation, cohesion, word order, and spelling), whose criteria are described in detail in Ogawa (2021). It turned out that approximately 84% of the errors were identified as mistranslation. In this present study, only the number of errors is discussed.

chance of contributing to excessive or structural zeros among all the zeros in the data. In this case, the ZI model tells us what affects the probability that a segment will have zero errors.

For this statistical analysis, packages called glmmTMB (Brooks et al., 2017) and DHARMa (Hartig, 2022) were used in RStudio (RStudio Team, 2022). A variable called Text, which identifies the six STs in CRITT TPR-DB, was used as a random effect for both count and ZI models. A backward step-wise selection method was adopted to build models; that is, a model was created with all the ST characteristics included in fixed effects, and one independent variable was removed at a time until all the fixed effects in the model were statistically significant ($p < 0.05$). Another set of models was created for translator characteristics in the same manner. The two different types of characteristics (ST and translator) were not combined in a single model so that the methodology would be identical to that of Ogawa (2021). The following is a sample model:

```
model <- glmmTMB(TotalErrorCount ∼ 1 + fixed1 + fixed2 + (1|Text),
                             zi=∼ fixed2 + (1|Text), data=df, family=”nbinom1”)
```

The independent variables tested in this study are described in Table 1 (see Ogawa 2021 for more detailed explanations of each variable). The first four are ST characteristics, and the last five are translator characteristics. Categorical variables are Figurative (3 levels), L1 (2 levels), InitialOrientation (4 levels), and EndRevision (3 levels). The rest are numeric variables.

Figurative	Refers to how many figurative expressions a segment contains.[4]
Ddepth	Refers to syntactic complexity of a segment. It counts the number of layers underneath the surface structure, processed by Berkeley Neural Parser (Kitaev et al., 2019; Kitaev and Klein, 2018). Higher values indicate greater syntactic complexity.
LWRatio	Refers to the proportion of words, per segment, that are longer than seven letters.
PROB1Norm	Refers to segment-level word frequency based on a log10 probability of a monogram ST word frequency calculated using the BNC corpus as a reference (Carl et al., 2016). The higher the value is, the greater the number of less common words a translator encounters in a segment.
Training	Indicates how many years of formal translation training a participant had.
Experience	Indicates how many years of translation experience a participant had.
L1	Indicates the participants’ first language, either Japanese or English.
Initial Orientation	Categorizes sessions into four groups depending on how long the participant read the ST before starting to produce their translation (see Dragsted and Carl, 2013): Head-starters (who immediately started typing), Quick-planners (who read the first few ST sentences before typing), Scanners (who quickly scanned through the ST), and Systemic-planners (who read the entire ST).
EndRevision	Classifies sessions into three categories depending on how much time the participant spent re-reading the ST or TT after completing a draft: Long (more than 25% of the session duration was used for revision), Short (less than 25% of the session duration was used), and None (end revision was not conducted).

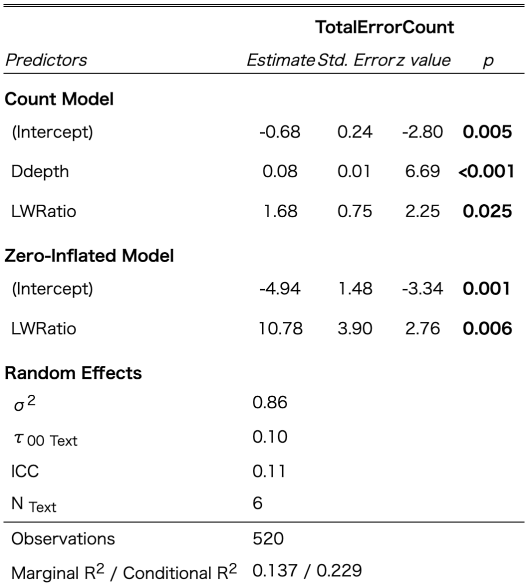
Table 1: Descriptions of ST/Translator Characteristics Used as Independent Variables

[4] This annotation has been revised and is therefore different from the annotation employed in Ogawa (2021), in which Figurative was a dichotomous annotation referring to whether a segment contains a metaphoric expression.

3 Results

3.1 Errors and ST Characteristics

The best model for estimating the number of errors from ST characteristics included Ddepth and LWRatio in the count model and LWRatio in the ZI model, and is summarized in Table 2.[5] The count model portion indicates that Ddepth and LWRatio positively impacted the TotalErrorCount. That is, the more syntactically complex the segment was and the greater number of long words the segment had, the greater number of errors the segment contained. Figure 2 visualizes a prediction based on this result, which shows that the predicted number of errors increases as Ddepth increases.[6] This tendency is maintained across different LWRatio values, and a greater number of errors are expected when LWRatio is higher.



	TotalErrorCount			
Predictors	*Estimate*	*Std. Error*	*z value*	*p*
Count Model				
(Intercept)	-0.68	0.24	-2.80	**0.005**
Ddepth	0.08	0.01	6.69	**<0.001**
LWRatio	1.68	0.75	2.25	**0.025**
Zero-Inflated Model				
(Intercept)	-4.94	1.48	-3.34	**0.001**
LWRatio	10.78	3.90	2.76	**0.006**
Random Effects				
σ^2	0.86			
$\tau_{00\ \text{Text}}$	0.10			
ICC	0.11			
N_{Text}	6			
Observations	520			
Marginal R^2 / Conditional R^2	0.137 / 0.229			

Table 2: Model Summary for ST Characteristics

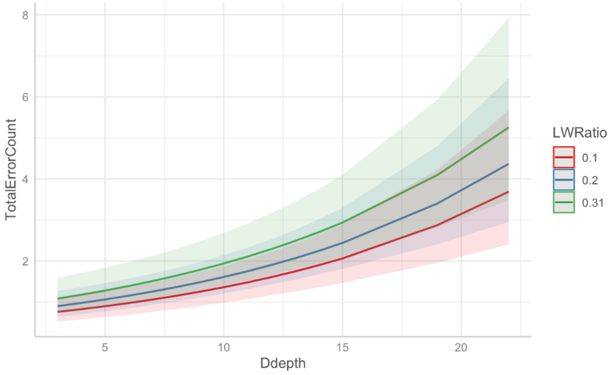


Figure 2: Predicted Number of Errors based on ST Characteristics

The middle section of Table 2 ("Zero-Inflated Model"), which should be interpreted as logistic regression, shows that LWRatio had a significant effect on excessive zeros. The positive estimate value, which is a log odds, indicates that it was more likely for a segment to be a member of excessive zeros as LWR increased. That is, when LWRatio was higher, a segment was more likely to contain zero errors. Converting the log odds to a probability (i.e., the exponential of the log-odds divided by the exponential of the log-odds plus one) suggests that a one-unit increase in LWRatio increases the chance of excessive zeros by 99%.

This is a puzzling result. How can LWRatio increase the number of errors while also increasing the chance of having zero errors with such a high probability? It might be because many of the segments with high LWRatio values contain zero errors. As Figure 3 illustrates, the segments with high (> 0.4) LWRatio only exist in Text 5, where the number of errors is relatively low. The two segments at the high end of LWRatio mostly contain zero errors, as

[5] The model summary tables in this study were produced using the *sjPlot* package (Lüdecke, 2021).

[6] The visualizations of predicted number of errors were produced using the *ggeffects* package (Lüdecke, 2018).

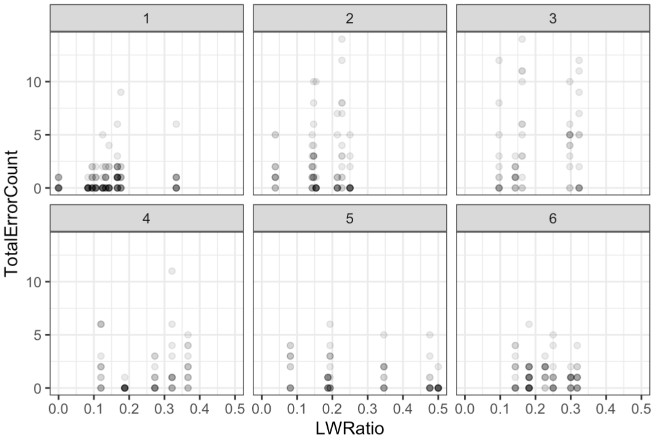


Figure 3: Distribution of LWRatio and TotalErrorCount in each ST

indicated by the dark dots. Also, the result might have been influenced by the fact that the range of LWRatio was too small (i.e., 0.0 to 0.5), which would make the change in log odds for a one-unit increase tremendous.

3.2 Errors and Translator Characteristics

The best model for estimating the number of errors based on translator characteristics included Experience, InitialOrientation and EndRevision in the count model and Training in the ZI model. Table 3 shows that, in the count model, Experience and EndRevision negatively influence the TotalErrorCount. That is, the participants who had more years of experience and spent more time on end revision made fewer errors. This is a somewhat expected (and pleasant) result.

As for InitialOrientation, Table 3 tells us that those who read STs before translating for at least some time (i.e., Quick-planners, Scanners, and Systemic-planners combined) made more errors than those who immediately started producing translation (i.e., Head-starters, the base level factor). This is a bit surprising given the fact that most errors in our dataset were mistranslation. Naively speaking, translators should be able to avoid making errors if they read the ST carefully, but this intuition was not supported by the result. Figure 4, which visualizes the predicted number of errors based on the count model, shows that Head-starters make the least number

	TotalErrorCount			
Predictors	*Estimate*	*Std. Error*	*z value*	*p*
Count Model				
(Intercept)	0.99	0.23	4.26	**<0.001**
Experience	-0.03	0.01	-3.77	**<0.001**
InitialOrientation [Quick-planner]	0.47	1.44	3.27	**0.001**
InitialOrientation [Scanner]	0.27	0.25	1.09	0.275
InitialOrientation [Systemic-planner]	0.35	0.17	2.09	**0.037**
EndRevision [Short]	-0.40	0.19	-2.13	**0.033**
EndRevision [Long]	-0.76	0.20	-3.74	**<0.001**
Zero-Inflated Model				
(Intercept)	-3.26	1.05	-3.11	**0.002**
Training	0.78	0.27	2.89	**0.004**
Random Effects				
σ^2	0.84			
$\tau_{00\ \text{Text}}$	0.17			
ICC	0.17			
N_{Text}	6			
Observations	492			
Marginal R^2 / Conditional R^2	0.145 / 0.287			

Table 3: Model Summary for Translator Characteristics

of errors, followed by Scanners, Quick-planners, and Systemic-planners in this order. Further examination revealed that the average years of experience per group decreased in the same order, although a statistically significant interaction effect was not found between InitialOrientation and Experience. It is also worth mentioning that InitialOrientation was annotated at the session level, not at the participant level, as some participants—regardless of their years of experience—spent very different amounts of time on ST reading across sessions. The result might have been different if the experiment had been conducted in a more ecologically valid situation, where the participants would exhibit their routine ST-reading habit.

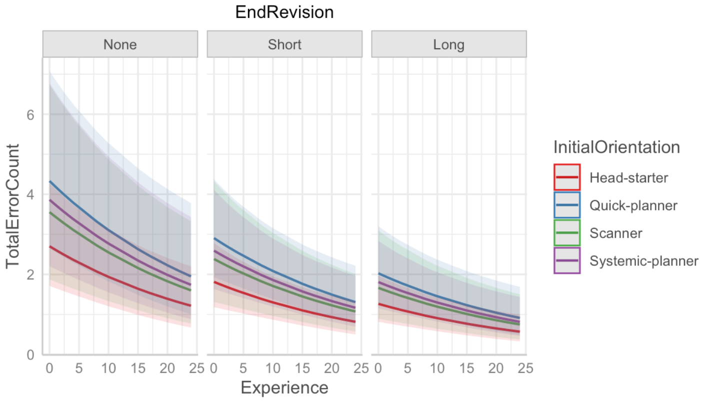


Figure 4: Predicted Number of Errors Based on Translator Characteristics

Figure 4 also makes it clear that the number of errors decreases as years of experience increases, and that the Long group in EndRevision (i.e., participants who spent more than a quarter of session duration on end revision) produces fewer errors than the other two categories. Even translators who have zero experience seem to be able to greatly reduce the number of errors by spending more time on end revision. Jakobsen (2002) found that professional translators spent a greater proportion of time on end revision than student translators, and he presumed that professionals monitored and optimized their draft to achieve higher quality. The present study corroborates his observation, providing evidence that longer end revision leads to fewer errors.

The ZI model in Table 3 indicates that Training had a positive impact. That is, the more training the participants had, the more likely a segment had zero errors. It is worth noting that Training was only significant in the ZI model. This suggests that years of training led to a significant difference in the production of error-free translation while it did not significantly reduce the number of errors. For example, a participant with one year of training is 69% (i.e., $exp(0.78)/(1 + exp(0.78))$) more likely to produce a translation with zero errors compared to a participant with no training, but when a participant with one year of training does make errors, the error count may or may not be lower than that of a translator with no training.

This result might suggest that, although training can help translators avoid making errors to a certain extent, having experience is crucial for overall translation quality. However, it may be necessary to consider what excessive zeros mean in the context of this specific translation experiment. In some studies, the concept is clear and easy to understand. For example, consider a situation where a researcher would like to know whether the number of visits to on-campus counseling services is influenced by the students' alcohol use. There would be many students who do not use counseling services at all, so the data would be zero-inflated. Among those zeros, students who are away from the campus or regularly see a counselor off campus would be members of excessive zeros because they are very unlikely to contribute to the count data. In our context, where participants translate English into Japanese without any external resources in an experimental setting, every participant can potentially make errors. Therefore, what excessive zeros are depends on how researchers interpret them.

Although this would greatly benefit from more discussion than we can achieve here, let us assume that excessive zeros represent an error-free translation produced when several factors coincide to create a “perfect situation,” where translators make no errors. This is only a hypothetical situation, as we do not know what exactly creates such a “perfect situation” for translators. Nonetheless, it is reasonable to assume that translators with high expertise are unlikely to make errors when translating a segment that is easy for them.[7] Interpreted this way, the present study suggests that years of training increase the chances a translator will be in one of these “perfect situations” wherein they make zero errors. Perhaps Shreve (2009), who suggests that translators can increase their level of expertise by developing metacognitive skills, can provide us with a potential explanation. For instance, if the participants in our dataset in fact underwent some training that improved their metacognitive skills and as a result acquired heightened awareness of what kind of errors they tend to make, the results of the ZI model can be interpreted as supporting evidence that such training does have a positive impact on translation quality.

4 Discussion

The check marks in Table 4 indicate which characteristics produced a significant effect on TotalErrorCount in this study. Dur, on the right, refers to production duration (i.e., the time taken to translate a given segment, including pauses), and the results shown here are from Ogawa (2021). Dur is utilized to quantify time efficiency here, so that it will be clear which ST and translator characteristics influence translation quality and time efficiency in parallel.

	TotalErrorCount (count)	TotalErrorCount (ZI)	Dur
Figurative			
Ddepth	✔		
LWRatio	✔	✔	
PROB1Norm			✔
Training		✔	
Experience	✔		✔
L1			
InitialOrientation	✔		
EndRevision	✔		✔

Table 4: Statistically Significant Characteristics

Figurative and L1 were not significant in any models. Figurative expressions have been discussed and identified as a source of translation difficulty (e.g., Schäffner, 2004; Sjørup, 2008). A preliminary analysis on TotalErrorCount indeed indicated that Figurative produced a significant result in the count model, though only if it was the sole fixed effect in a model. Using the backward step-wise selection method may have lowered the explanatory power of Figurative when other independent variables were involved in a model. This might also be true for L1, which showed a significant result in the ZI model in a single fixed-effect model.

There was no ST characteristic that significantly influenced both TotalErrorCount and Dur, but two translator characteristics (i.e., Experience and EndRevision) were important factors for both dependent variables. Ogawa (2021) revealed that Dur was negatively influenced by

[7] Note that ease/difficulty are necessarily subjective. A segment can be easy for a translator if it is embedded in a rich context in their familiar domain without any words that they do not know.

Experience and positively influenced by EndRevision. That is, more experienced translators translated faster, and participants who spent more time on end revision had longer production duration as a result.[8] Combined with the finding in the present study, it can be concluded that i) years of experience is a good predictor of translation quality and time efficiency and that ii) the time translators spend on end revision inevitably increases production duration but in return increases quality.

Dur was also significantly influenced by PROB1Norm; participants spent more time translating as they encountered a greater number of less familiar or less frequently used words. Of course, each individual has different linguistic knowledge, and PROB1Norm is a simplistic operationalization of word familiarity. That being said, if PROB1Norm truly impacts Dur but not TotalErrorCount, it might be the case that translators can improve time efficiency by further familiarizing themselves with the source language. Familiarization would particularly matter when it comes to different genres or domains, where words are used as terms with different meanings than when they are used in general texts.

Ddepth and LWRatio increased the error count while Experience and EndRevision decreased it, as discussed in the previous section. The former two are ST characteristics that are fairly easy to quantify. It is not clear at this point whether pre-editing STs in such a way that Ddepth and LWRatio values will be lower leads to higher-quality translation. These characteristics may nevertheless be used to compare different texts and/or caution translators about potential difficulty in advance. The effects of Experience and EndRevision were also straightforward. This evidence may encourage translators to gain more experience and keep in mind that the end revision phase is critical to translation quality even when translators feel the need to prioritize time.

It may be worth mentioning that no clear relationship was observed between EndRevision and Experience. Recall that, in InitialOrientation, the Head-starter group was expected to produce the least number of errors and that the Quick-planner group the most. This can be explained by the fact that these two groups had the highest and lowest average years of experience respectively. In contrast, the Short group in EndRevision had the highest average years of experience, followed by the Long and None groups in this order. Moreover, participants who did not spend any time reviewing their draft (i.e., the None group) were most prevalent in the Head-starter group, and none of them belonged to the Quick-planner group. Although there was no clear relationship between EndRevision and InitialOrientation in this study, previous research has found that Head-starters and Quick-planners tended to prefer online revisions (i.e., revising as they produce a TT) while Scanners and Systemic-planners carried out end revision (Dragsted and Carl 2013). Perhaps, TotalErrorCount may be better analyzed if participant revision preferences, including online revision as well as end revision, are taken into consideration.

5 Future Directions

This paper has revealed that some ST and translator characteristics significantly contribute to the number of errors per segment in English-Japanese from-scratch translation. Combined with findings from a previous study, evidence was found that translators' years of experience make a difference in terms of translation quality and time efficiency, and that the length of end revision has a positive effect on quality even though it may take some extra time. However, the examination of translators' initial orientation phase with a ST suggested that there may be a complex interplay between the length of end revision and translator style (i.e., translator preferences for revisions and initial ST reading). This should be further scrutinized in future research.

[8] Note that EndRevision was defined by gaze data, not by typing activity. However, the fact that Dur was positively influenced by EndRevision may suggest that many of the participants who conducted end revision ended up making changes to their original draft.

This study was limited to English-Japanese translation, and hence, the result should be corroborated by similar studies using other languages. In doing so, methodology needs to be discussed in two respects. Firstly, the way of quantifying translation quality is of utmost importance since it can produce very different results. Error count is relatively easy to use as a quantification of translation quality but fails to recognize fine-grained differences in quality (e.g., it does not distinguish excellent from adequate quality). Some researchers have tried evaluating the quality of HT using metrics primarily used for machine translation (MT) output, such as BLEU (e.g., Carl and Buch-Kromann, 2010), and produced interesting results. At the same time, however, research has found that those metrics cannot fully capture errors in HT because HT errors are different from MT errors (Specia and Shah, 2014). Translation evaluation methods call for further discussions in TPR.

Even if the number of errors is used as a primary measure of translation quality, multiple evaluators and calculations of inter-rater reliability may need to be considered. Our study was limited to errors annotated by a single researcher, which admittedly is the biggest weakness of this paper. Furthermore, there may be a better way of counting errors. The method utilized (see Section 2) seems justifiable since it allows us to quantify errors regardless of the target language and to compare different studies in the CRITT TPR-DB. However, it does require significant manual work and may not be viable when multiple evaluators are involved.

The other methodological factor that demands attention is the use of ZIGLMMs. This is a fine-tuned statistical method that can deal with zero-inflated count data, but the interpretation of ZI models requires more discussion in TPR. Some researchers may find it implausible to assume excessive zeros in conducting this line of analysis.

Since many researchers in TPR use linear mixed effect models, it might be time for us to discuss what is considered a good model in our discipline. In this paper, the R^2 of the best model was 0.29, which means that roughly 30% of the total variance was explained by the model. This seems to be satisfactory as much smaller numbers have been reported (e.g., Ogawa, 2021; Vanroy et al., 2021), while much greater values have been achieved as well (e.g., Heilmann and Llorca-Bofí, 2021). Of course, it is risky to solely rely on R^2 as if it were the only criteria that could be used to validate an analysis. Such a discussion will surely lead to the advancement of methodology in TPR.

References

Brooks, M. E., Kristensen, K., van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., Skaug, H. J., Maechler, M., and Bolker, B. M. (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal*, 9(2):378–400.

Carl, M. and Buch-Kromann, M. (2010). Correlating translation product and translation process data of professional and student translators. In *Proceedings of the 14th Annual conference of the European Association for Machine Translation*, Saint Raphaël, France. European Association for Machine Translation.

Carl, M., Schaeffer, M., and Bangalore, S. (2016). The CRITT translation process research database. In *New directions in empirical translation process research*, pages 13–54. Springer.

Colina, S. (2009). Further evidence for a functionalist approach to translation quality evaluation. *Target. International Journal of Translation Studies*, 21(2):235–264.

Daems, J., Carl, M., Vandepitte, S., Hartsuiker, R., and Macken, L. (2016). The effectiveness of consulting external resources during translation and post-editing of general text types. In *New directions in empirical translation process research*, pages 111–133. Springer.

Daems, J., Macken, L., and Vandepitte, S. (2013). 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 2nd Workshop on Post-editing Technology and Practice*.

Dragsted, B. (2005). Segmentation in translation: Differences across levels of expertise and difficulty. *Target-international Journal of Translation Studies*, 17(1):49–70.

Dragsted, B. and Carl, M. (2013). Towards a classification of translation styles based on eye-tracking and key-logging data. *Journal of Writing Research*, 5(1):133–157.

Fields, P., Hague, D. R., Koby, G. S., Lommel, A., and Melby, A. (2014). What is quality? A management discipline and the translation industry get acquainted. *Revista Tradumàtica: tecnologies de la traducció*, 12:404–412.

Garvin, D. A. (1984). What does product-quality really mean. *Sloan management review*, 25:25–43.

Germann, U. (2008). 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*, pages 20–23.

Hartig, F. (2022). *DHARMa: Residual Diagnostics for Hierarchical (Multi-Level / Mixed) Regression Models*. R package version 0.4.5.

Heilmann, A. and Llorca-Bofí, C. (2021). Analyzing the effects of lexical cognates on translation properties: A multivariate product and process based approach. In *Explorations in Empirical Translation Process Research*, pages 203–229. Springer.

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

Jakobsen, A. L. (2002). Translation drafting by professional translators and by translation students. *Copenhagen studies in language*, 27:191–204.

Jakobsen, A. L. and Jensen, K. T. H. (2008). Eye movement behaviour across four different types of reading task. *Copenhagen studies in language*, 36:103–124.

Kitaev, N., Cao, S., and Klein, D. (2019). Multilingual constituency parsing with self-attention and pre-training. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3499–3505, Florence, Italy. Association for Computational Linguistics.

Kitaev, N. and Klein, D. (2018). Constituency parsing with a self-attentive encoder. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2676–2686, Melbourne, Australia. Association for Computational Linguistics.

Kivilehto, M. and Salmi, L. (2017). Assessing assessment: The authorized translator's examination in finland. *Linguistica Antverpiensia, New Series–Themes in Translation Studies*, 16:57–70.

Lommel, A. (2018). Metrics for translation quality assessment: a case for standardising error typologies. In *Translation Quality Assessment*, pages 109–127. Springer.

Lüdecke, D. (2021). *sjPlot: Data Visualization for Statistics in Social Science*. R package version 2.8.7.

Lüdecke, D. (2018). ggeffects: Tidy data frames of marginal effects from regression models. *Journal of Open Source Software*, 3(26):772.

Mossop, B. (2014). *Revising and editing for translators*. Routledge.

Ogawa, H. (2021). *Difficulty in English-Japanese Translation: Cognitive Effort and Text/Translator Characteristics*. Doctoral thesis, Kent State University.

O'Brien, S. (2012). Towards a dynamic quality evaluation model for translation. *The Journal of Specialised Translation*, 17(1):55–77.

RStudio Team (2022). *RStudio: Integrated Development Environment for R*. RStudio, PBC, Boston, MA.

Schäffner, C. (2004). Metaphor and translation: Some implications of a cognitive approach. *Journal of pragmatics*, 36(7):1253–1269.

Sharmin, S., Spakov, O., Räihä, K.-J., and Lykke Jakobsen, A. (2008). Where on the screen do translation students look while translating, and for how long? *Copenhagen Studies in Language*, 36:31–51. Looking at Eyes: Eye-Tracking Studies of Reading and Translation Processing. (red.) Susanne Göpferich; Arnt Lykke Jakobsen; Inger M. Mees.

Shreve, G. M. (2009). Recipient-orientation and metacognition in the translation process. In Dimitriu, R. and Shlesinger, M., editors, *Translators and their readers: in homage to Eugene A. Nida*, pages 257–270. Les Editions du Hazard, Brussels.

Sjørup, A. C. (2008). Metaphor comprehension in translation: Methodological issues in a pilot study. *Copenhagen studies in language*, 36:53–77.

Specia, L. and Shah, K. (2014). Predicting human translation quality. In *Proceedings of the 11th Conference of the Association for Machine Translation in the Americas: MT Researchers Track*, pages 288–300.

Vanroy, B. (2021). *Syntactic Difficulties in Translation*. Doctoral thesis, Ghent University.

Vanroy, B., Schaeffer, M., and Macken, L. (2021). Comparing the effect of product-based metrics on the translation process. *Frontiers in Psychology*, 12.