Translation Quality in Post-Edited versus Human-Translated Segments: A Case Study

Elaine O'Curran elaine.ocurran@welocalize.com
Welocalize Inc., 6 Dundee Park, Andover, Massachusetts 01810, USA

Abstract

We analyze the linguistic quality results for a post-editing productivity test that contains a 3:1 ratio of post-edited segments versus human-translated segments, in order to assess if there is a difference in the final translation quality of each segment type and also to investigate the type of errors that are found in each segment type. Overall, we find that the human-translated segments contain more errors per word than the post-edited segments and although the error categories logged are similar across the two segment types, the most notable difference is that the number of stylistic errors in the human translations is 3 times higher than in the post-edited translations.

1. Introduction

We have come to expect that using machine translation (MT) in combination with human post-editing (MTPE) saves on time and cost when compared with human translation (HT). However, there remains a lot of fear in the industry that integrating MT into the translation workflow will lead to lower quality. In order to find out for ourselves, we performed a detailed analysis of the linguistic quality assessment (LQA) reports for a post-editing productivity test we had recently carried out that involved the languages Brazilian Portuguese, French and Spanish and 6 translators (two per locale).

In some cases at the request of a client or an internal team, we conduct productivity tests in order to evaluate the usability of the raw MT output for post-editing by human translators and to estimate productivity gains over human translation from scratch. These tests are carried out in iOmegaT[1], an instrumented version of the open source translation tool OmegaT. Productivity tests are typically carried out for 8 hours and involve 2 translators per local. Of these 8 hours, 1 hour is used for revision. This is an important step for translators, as it mirrors their usual approach. As a testing environment, iOmegaT minimizes the disruption to a post-editor's process by providing an interface similar to those most frequently used in everyday production. The translation kit contains a mix of MT and HT segments as well as the usual formatting features (e.g. tags), and MT segments can be distinguished from internally propagated fuzzy matches. However, the kits are usually prepared to contain as few fuzzies as possible in order to maximize the MT and HT words in the tests. Glossaries and translation memories can be provided in the tool, translators can revisit segments and there is a spellcheck and tag validation feature for the review phase. Following delivery, a test kit is sent for LQA to assess the quality against the prescribed benchmarks for the content type, domain and language. By running this additional check, we can be certain that productivity gains are valid, not occurring at the expense of quality.

[1] CNGL Invention Disclosure, option period triggered 22/06/2012

2. Related Work

Fiederer and O'Brien (2009) set out to investigate if post-edited MT output was necessarily of lower quality than human translation and found that the post-edited machine translated output was assessed to be of higher clarity and accuracy, while the human translations were assessed to be of better style. Findings by Guerberof (2010) suggest that translators produce higher quality when using machine translated output than when processing fuzzy matches from translation memories. In her experiment, the number of errors found in TM segments was 91% higher than in MT segments. MT segments, on the other hand, contained 26% more errors than in the HT segments.

Plitt & Masselot (2010) used 12 professional localisation translators in their study and reported that translation jobs contained a higher number of errors than post-editing jobs. The MT engine had been trained on a large amount of company data, as this was a study carried out by Autodesk, and this scenario is highly representative for our case study.

Garcia (2010) was surprised at the quality results in his comparative study, which showed that the MT passages populated by Google Translator Toolkit and subsequently human-edited were more favourably assessed by the reviewers in 33 of 56 cases. This suggests that translating by post-editing MT output may be advantageous (Garcia, 2010). However, it is important to note that Garcia used trainee translators is in his study, whereas all the others studies referenced here employed professionals.

From her analysis of several months' LQA data for 4 language pairs, gathered from both human-translated and post-edited content, Peruzzi (2013) concludes that "the main differences in quality and types of errors are found between languages rather than translation scenarios, and […] these differences may not only be caused by quality of MT, but also by different cultural and linguistic aspects" (2013). Peruzzi's evaluation was based on two different workflows – one including MTPE and one including HT. The current use case differs in that it is based on a workflow that contains a mix of MTPE and HT segments within the same test environment.

3. Methods

3.1. MT profile

The machine translations are provided by a statistical MT system that has been specially customized for the client content using translation memories and glossaries.

3.2. Translator profile

All six translators are familiar with the account that is being tested and have at least 5 years translation experience and with the exception of the two, very senior, Spanish resources, all have between 1 to 4 years of post-editing experience.

3.3. Content profile

The content translated and post-edited in the test kits is real User Assistance content from the Software Antivirus/Security Compliance domain.

3.4. Reviewer profile

Dedicated third party account reviewers performed the LQA on the productivity kits to check compliance with standard quality expectations for the account. It was a blind review, i.e. the reviewers were not aware if the segment origin was MTPE or HT.

3.5. Linguistic review method

Similar to the LISA QA Model and SAE J2450, our applied QA metrics are a quantitative-based method of translation quality assessment which measures the number, severity and type of errors found in a text and calculates a score based on the number of words reviewed, indicative of the quality of a given translation. The reviewers evaluate the translations based on the following criteria:

1. Accuracy: Cross References, Omission/Addition, Incorrect Translation/Meaning, Unlocalized Text
2. Language: Punctuation, Spelling/Typo, Grammar/Syntax
3. Terminology: Context, Inconsistency, Glossary
4. Style: General Style, Client Style Guide, Language Variants/Slang, Register/Tone, Unnecessary Additions
5. Country: Country/Regional Standards, Local Market Suitability
6. Functional: Format, Hidden Text, Tags/Links, Technical Procedures, Spacing

The severity levels Minor or Major are applied to each error, based on the definitions in Table 1 below.

Major errors are blatant and severe errors that jeopardize, inverse or distort the meaning of a translation. Major errors are severe failures in accuracy, compliance, or language. Examples:
– Any statement that can be potentially offensive.
– Errors that endanger the integrity of data or the health/safety of users.
– Errors that modify or misrepresent the functionality of the device or product.
– Errors that clearly show that the client's and/or Welocalize' instructions haven't been followed.
– Errors that appear in a High Visibility Portion and/or is numerously repeated.
– Grammar or syntax errors that are gross violations of generally accepted language conventions.
Minor errors are all errors that do not fall under major severity as defined above nor are merely preferential changes. Examples:
– Accuracy errors that result in a slight change in meaning.
– Small errors that would not confuse or mislead a user but could be noticed.
– Formatting errors not resulting in a loss of meaning, e.g. wrong use of bold or italics.
– Wrong use of punctuation or capitalization not resulting in a loss of meaning.
– Generic error to indicate generally inadequate style (e.g. literal translation, "stilted" style, etc.).
– Grammar or syntax errors that are minor violations of generally accepted language conventions.
– Typos and misspellings that do not result in a loss of meaning.

Table 1: Error Severity Descriptions

3.6. LQA results

For this productivity test, 3 of the kits returned a fail. This was one of the drivers behind this case study, as we wanted to understand if the underediting in these kits could be traced back to MT segments that had not been post-edited properly, or if the translators had simply not performed an adequate self-review in general. It is worth noting that the two inexperienced post-editors delivered the best overall quality.

Translator	LQA Result
ES - 1	PASS
ES - 2	PASS
FR - 1	PASS
FR - 2	FAIL
PTBR - 1	FAIL
PTBR - 2	FAIL

Table 2: LQA results per resource

The Pass treshold is 99.60% based on the following mathematical algorithm : =(1-(Minor_Errors+(2*Major_Errors*Major_Errors))/Sample_Size).

3.7. Review scope

Across three locales and 6 translators, approximately 8000 words were reviewed. Figure 1 below illustrates the exact number of words that were reviewed per locale and per segment type. The reason we see less words for Brazilian Portuguese is that the reviewers were instructed to spend 1 hour on each translator's kit and mark all the segments that had been checked. Due to the higher number of errors for this locale, the reviewer covered less words.

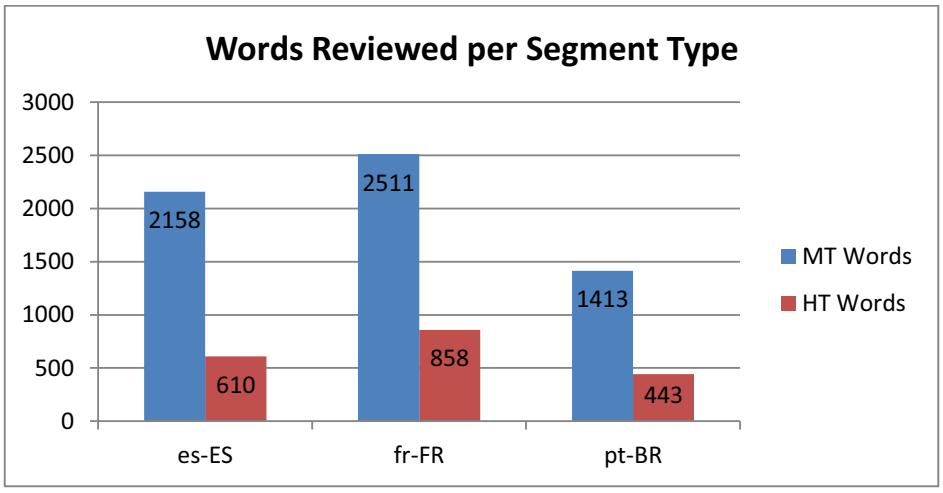

Figure 1: Words reviewed by locale and segment type

Figure 1 also illustrates the approximate 3:1 ratio of MT versus HT words in the reviewed kits.

4. Results

4.1. Errors per 1000 words

To account for the difference in word count per segment type, we calculate errors per 1000 words. As illustrated in Figure 2, we found that there were more errors in HT segments than in MT segments across all three locales. This consistent result was surprising considering the different levels of quality that had been delivered for this test.

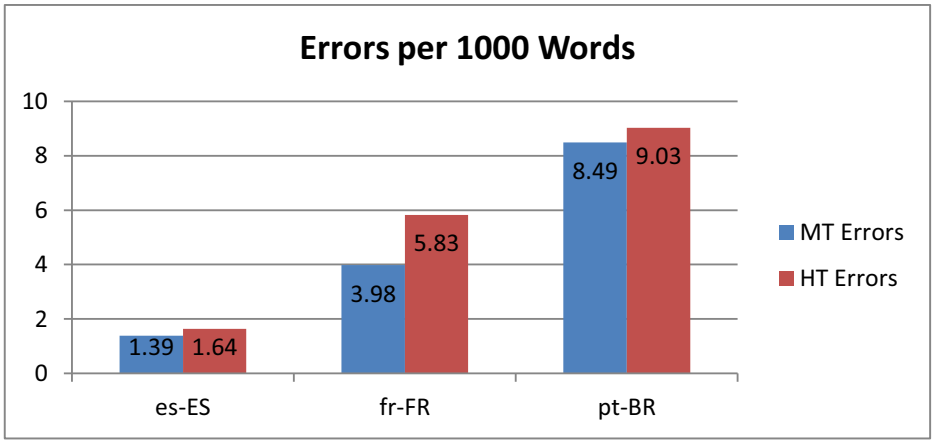

Figure 2: Errors per 1000 words

4.2. Error types found

For illustration purposes, the Table 3 error table below has been normalized to account for the 3:1 ratio of MTPE words versus HT words. The HT errors have simply been multiplied by three. We see that overall there are 25 errors in MTPE, including 5 Major errors, while there are 30 errors in HT including 6 Major errors. There are more Language, Style and Tag errors in HT, while MTPE has more Glossary errors. The Accuracy errors are quite evenly distibuted: both have exactly 6 Minor and 3 Major errors in the Accuracy category. The most notable difference is that the number of stylistic errors in the HT segments is 3 times higher than in the MTPE segments.

Error Category	MTPE Errors	HT Errors
Accuracy - Meaning/Incorrect Translation - Major	1	0
Accuracy - Meaning/Incorrect Translation - Minor	3	6
Accuracy - Omissions/Additions - Major	2	3
Accuracy - Omissions/Additions - Minor	3	0
Functional - Tags/Links - Major	1	3
Functional - Tags/Links - Minor	1	0
Language - Grammar/Syntax - Major	1	0
Language - Grammar/Syntax - Minor	6	6

Language - Punctuation - Minor	0	3
Style – Client Style guide - Minor	1	0
Style - Language variants/slang - Minor	1	0
Style - General style - Minor	0	3
Style - Unnecessary Additions - Minor	0	3
Terminology - Glossary adherence - Minor	5	3
Total errors	**25 (5 major)**	**30 (6 major)**

Table 3: Error types in MTPE and HT

5. Conclusions

The result of this case study supports the findings of some of the other studies mentioned above that found fewer errors in MT post-edited work than in human translations. While there is some overlap between the types of errors found in the human-translated segments and post-edited segments, it is notable that more errors were found in human translations in categories such as Punctuation, Tags and Style. However, our case study was performed on relatively small volumes, the three locales are Romance languages and the content type is technical. In order to draw firm conclusions, it would be important to conduct a larger study with more diverse languages and content types and to also include fuzzy matching for additional benchmarking.

Acknowledgement

The author would like to acknowledge the support of Olga Beregovaya, Laura Casanellas and Lena Marg.

References

Fiederer, R. and O'Brien, S. (2009) Quality and Machine Translation: A realistic objective? *The Journal of Specialised Translation*, Issue 11. http://www.jostrans.org/issue11/art_fiederer_obrien.pdf. Accessed 17 July 2014

Garcia, I. (2010) Is machine translation ready yet? *Target*, 22(1): 7–21

Guerberof, A. (2009) Productivity and Quality in MT Post-Editing, Universitat Rovira I Virgilli, Spain. http://www.mt-archive.info/MTS-2009-Guerberof.pdf Accessed 17 July 2014.

Peruzzi, S. (2013) Investigating the Impact of Machine Translation and Post-Editing on Quality and Errors in a Translation Memory-based Workflow. Dissertation for the MSc in Translation Technology at DCU, Ireland

Plitt, M., and Masselot, F. (2010) A Productivity Test of Statistical machine Translation Post Editing in a Typical Localisation Context. *The Prague Bulletin of Mathematical Linguistics*. Jan. Num 93. 7-16.