Machine Translation and Post-editing: Impact of Training and Directionality on Quality and Productivity

M. Cristina Toledo Báez

M. Cristina Toledo Báez
University of Málaga
toledo@uma.es;
ORCID:
0000-0003-0604-797X

Abstract

The aim of this pilot study is two-fold: on the one hand, it is to explore the impact of participants' educational background in both post-editing quality and productivity and, on the other, to compare quality and speed of post-editing into L1 with quality and speed of post-editing into L2.

Keywords: post-editing, quality, productivity, training, directionality, machine translation.

Resum

L'objectiu d'aquest estudi és doble: per una banda, explorar la influència que té la formació dels participants tant en la qualitat com en la productivitat de la postedició; per l'altra, comprar la qualitat i la velocitat de postedició en la L1 i la qualitat i la velocitat de la postedició en la L2.

Paraules clau: postedició, qualitat, productivitat, formació, direccionalitat, traducció automàtica.

Resumen

El objetivo de nuestro estudio piloto es doble: por un lado, explorar el impacto de la formación de los participantes tanto en la calidad como en la productividad de la posedición y, por otro, comparar la calidad y la velocidad al poseditar a la L1 frente a la calidad y velocidad al posteditar a la L2.

Palabras clave: posedición, calidad, productividad, formación, direccionalidad, traducción automática.

Rebuda: 17 d'abril de 2018 / Acceptació: 29 de novembre de 2018 / Publicació: 18 de desembre de 2018

1. Introduction and related work

Post-editing has become a common practice within companies since it implies an increase in productivity when compared to human translation (Aranberri et al., 2014) without having a negative impact on quality (Plitt and Masselot, 2010).

Two elements of post-editing are of paramount importance for our pilot study: directionality and training. With regard to directionality, previous empirical studies have proved that post-editing into L2 presents similar results to post-editing into L1. Sánchez-Gijón and Torres-Hostench (2014) compared native and non-native speakers and they found out that the results for the best participants from group A (non-native) were very similar to those of group B (native) in a "good enough" post-editing task. In a study of post-editing effort in the Chinese-Portuguese language pair, Igor et al. (2017) reported no significant differences in directionality. There are even studies such as Garcia's (2011) that have demonstrated that translation trainees post-editing into L2 worked better (i.e. obtained higher marks) than post-editing into L1.

Regarding training, the literature has explored different aspects such as language proficiency and its relationship to training (Yamada, 2014), the integration of machine translation workflow into the curriculum (Kenny and Doherty, 2014), the use of tailored post-editing guidelines (Flanagan and Christensen, 2014) the creation of post-editing courses across the translation curriculum (Mellinger, 2017). Other key aspects in training are, on the one hand, the need for specific training (Flanagan and Christensen, 2014, among others) and, on the other, the target group to become post-editors. Translators have traditionally been the main target group since they have expert knowledge in source and target language and are familiar with target audience expectations (O'Brien, 2002). Nevertheless, some authors, such as Temizöz (2013), have explored post-editor profiles with different educational backgrounds. Her study compared the differences in post-editing performance between engineers and professional translators and the author concluded that results were similar in terms of quality.

2. Participants, aims and research questions

The study presented in this paper combines two elements – training and directionality – in an empirical pilot study conducted in 2015 on 10 undergraduate students enrolled in the 'Machine Translation and Post-Editing' course in the Bachelor's Degree in Modern Languages and Translation at the University of Alcalá (Spain). The participants presented two different profiles:

1. Two educational profiles: 1) Group A: five Modern Languages students without training in translation as they were Socrates/Erasmus exchange students from Bachelor's Degrees in Modern Languages and 2) Group B: five Translation students from University of Alcalá.
2. Two language profiles: 1) Group A: the five Modern Languages students were native English speakers and 2) Group B: the five Translation students were

native Spanish speakers. A language test conducted in class before the experiments yielded the following results: 1) Group A: all students had C2 level in English; 3 of them had B2 level and 2 of them C1 level in Spanish; 2) Group B: all students had C2 level in Spanish; 2 of them had B2 level and 3 of them C1 level in English.

The aims of our study are the following:

1. To explore the impact of participants' training on both post-editing quality and productivity. Our study aims at testing whether groups A and B would present different post-editing results regarding quality and productivity after having completed both groups the 'Machine Translation and Post-Editing' course.
2. To compare quality and speed of post-editing into L1 with quality and speed of post-editing into L2. Our study aims at testing whether groups A and B would present different post-editing results when directionality of post-editing is involved.

We will try to answer the following research questions:

1. To what extent is the post-editing performed by group A (Modern Languages students) linguistically correct and accurate?
2. To what extent is the post-editing performed by group B (Translation students) linguistically correct and accurate?
3. To what extent is the post-editing performed into L1 by both groups linguistically correct and accurate?
4. To what extent is the post-editing performed into L2 by both groups linguistically correct and accurate?
5. Is productivity (i.e., post-editing speed) influenced by training?
6. Is productivity (i.e., post-editing speed) influenced by directionality?

3. Materials

Three materials have been used in our study: 1) A template for evaluating the post-editing results based on TAUS error typology guidelines, 2) TAUS Quality Dashboard to register post-editors' productivity and 3) a corpus of machine-translated texts to be post-edited and post-edited texts.

3.1. Template for evaluating the post-editing results

The template to evaluate the quality of post-edited texts is based on the TAUS Error typology guidelines (TAUS, 2013). The TAUS typology has five main categories:

accuracy, language, terminology, style and country standards. The categories chosen for our template are two: accuracy and language. We disregarded the style category because it is not relevant for our study. Terminology was also disregarded since it should only be applied when a glossary or terminology source is provided (TAUS, 2013). As the texts to be post-edited in our study do not contain country standards, this category has also been disregarded.

Regarding accuracy subcategories, only one has been selected: mistranslation. The six other subcategories are not applicable since those errors are not found in the machine translated texts used in our corpus (see 3.3.). With regard to language, we have decided to make use of the three subcategories presented in TAUS error typology: grammar-syntax, punctuation, and spelling. Our adapted template for evaluating the post-edited results is presented in Table 1:

Accuracy	Incorrect interpretation of source text- mistranslation
Language	Grammar –syntax: non-compliance with target language rules Punctuation: non-compliance with target language rules Spelling: errors, accents, capital letters

Table 1: Error typology used for evaluating post-editing results

3.2. TAUS Quality Dashboard

TAUS Quality Dashboard has been used for two purposes: 1) the environment where the post-editors post-edited the machine translated texts and 2) the environment where post-editors' productivity is registered. Figure 1 shows a sample of the post-editing activity.

Figure 1: Sample of post-editing activity on TAUS Quality Dashboard

Productivity is measured in terms of processing speed, i.e. the time spent by each post-editor to edit an average number of words processed in a given timespan. Despite existing different methods to measure productivity such as, for instance, human-targeted translation edit rate (HTER) (Snover et al., 2006), as O'Brien (2011) pointed out, improving processing speed is the primary interest of the translation industry and post-editor trainees need to be familiar with this requirement.

3.3. Corpus

The corpus of machine translated texts and post-edited texts is taken from the CRITT Translation Process Database[1] from the Center for Research and Innovation in Translation and Translation Technology (Copenhagen Business School). Specifically, two pieces of news were chosen: Text 1, from BML12 study[2], a 132-word text in English, and Text 2, from GS12 study, a 169-word text in Spanish. Both texts were automatically translated with the hybrid system (statistical and rule-based machine translation) Google Translate and then post-edited on TAUS Quality Dashboard.

[1] <https://sites.google.com/site/centretranslationinnovation/tpr-db>
[2] BML12 and GSL12 studies can be accessed at <http://dighum1.ftsk.uni-mainz.de/cgi-bin/yawat/yawat.cgi> with the user TPRDB and the password tprdb.

The corpus used in the study is a parallel corpus encompassing 20 texts: 10 post-edited versions of Text 1 (5 post-edited versions by group A and 5 post-edited versions by group B) and 10 post-edited versions of Text 2 (5 post-edited versions by group A and 5 post-edited versions by group B).

4. Method and procedure

The study was conducted in the 'Machine Translation and Post-Editing' course of the Bachelor's Degree in Modern Languages and Translation at the University of Alcalá (Spain) in 2015/2016. This 8 ECTS credit course is elective for 3rd and 4th year students. The aim of this course is to provide a theoretical and practical approach to Machine Translation (MT) and Post-editing (PE). Three full-post-editing hands-on practice activities on TAUS Quality Dashboard are included in the course.

In terms of procedure, the pilot study took place in the same laboratory where the 'Machine Translation and Post-Editing' course had been taught. The empirical pilot study was conducted on 16 December 2015, the last scheduled class. Before starting the pilot study, students filled and signed a consent-to-participate form and they were explained the aims of the study.

After the explanation, they were asked to full-post-edit (i.e. to reach a quality similar to high quality human translation) Text 1 and Text 2. No time limit was set. The order for post-editing was the following: both groups post-edited Text 1, there was a 15-minute break and then both groups post-edited Text 2.

With regard to the evaluation process, the post-edited segments were assessed anonymously by an external evaluator, a Translation and Interpreting lecturer with experience in evaluating post-editing with TAUS template. Unfortunately, the addition of *a second evaluator was* not possible for logistical reasons. Corrected edited segments were counted as successful edits; segments in which the error had not been detected were counted as unsuccessful edits. Preferential errors (i.e. items that are not wrong per se, but where another solution is preferable) were also counted as successful edits.

5. Results

The results presented in this section were calculated with R, the free software environment for statistical computing, version 3.5.0. A one-way parametric analysis of variance (ANOVA) between subjects was conducted to test whether the differences between groups A and B were statistically significant ($p>.05$). The Shapiro-Wilk test showed that the data was normally distributed, and the Bartlett's test revealed that the assumption of homogeneity of variance for the analysis of variance was not violated ($p>.05$).

5.1. Results for Text 1 (post-editing from English into Spanish)

Examples of errors found in Text 1 are wrong grammar mood ("se prevé que avergüenza"), comma separating subject and verb ("China, debería"), wrong use of capital letters ("Gobierno chino") and very literal translation resulting in mistranslation ("contra la quema").

Table 2 below shows the arithmetic means of successful edits (AM) along with the ANOVA results for the degrees of freedom (df), the F value (F) and the Sig. value for Text 1.

Subcategory	AM for group A	AM for group B	df	F	Sig
Grammar/syntax	55.28%	74.52%	1	4.791	.06
Punctuation	40%	68%	1	5.444	.0479
Spelling	40%	80%	1	5.12	.0535
Mistranslation	37.5%	75%	1	18	.00283

Table 2. Arithmetic means of successful edits and ANOVA results for Text 1

Results show that group B obtained more successful edits than group A in all subcategories, albeit with the only significant different at the $p<.05$ in punctuation ($p=.0479$) and mistranslation ($p=.00283$).

5.2. Results for Text 2 (post-editing from Spanish into English)

Examples of errors found in Text 2 are wrong syntax ("it was not present" instead of "it had no prior presence"), wrong use of commas ("knowledge, and now it"), incorrect use of the plural form ("500 millions users"), mistranslation ("Google made a splash in the pool" as translation for "Google dio un golpe en la mesa"). Table 3 below shows the results for Text 2.

Subcategory	AM for group A	AM for group B	df	F	Sig
Grammar/syntax	69.98%	63.3%	1	0.335	.579
Punctuation	62.82%	59.96%	1	0.067	.803
Spelling	46%	60%	1	1.849	.211
Mistranslation	50.18%	58.62%	1	1.005	.345

Table 3. Arithmetic means of successful edits and ANOVA results for Text 2

Results show that group A obtained more successful edits than group B in grammar/syntax and punctuation subcategories whereas group B obtained more

successful edits than group A in spelling and mistranslation subcategories. However, the difference between both groups is not statistically significant in any subcategory.

5.3. Productivity results

Table 4 below shows the results for productivity in terms of minutes spent in post-editing in each text along with the ANOVA results.

Text	AM for group A	AM for group B	df	F	Sig
Text 1	15.8	13.6	1	2.017	.193
Text 2	14.6	14.8	1	0.04	.846

Table 4. Arithmetic means of minutes and ANOVA results for productivity

Results show that group B post-edited faster than group A, but the difference between both groups is not statistically significant.

6. Conclusions and future work

After having conducted our pilot study, we are able to answer the research questions raised above. Research questions number 1 and 2 were related to the impact of training on the linguistically correctness and accuracy of post-editing. Specifically, they sought to answer to what extent post-editing performed by group A (question 1) and by group B (question 2) is linguistically correct and accurate. Results showed that group A (Modern Languages students) only obtained more successful edits than group B (Translation students with training on Translation) in the grammar/syntax and punctuation subcategories in Text 2. Group B obtained more successful edits than group A in all subcategories in Text 1 and in the spelling and mistranslation subcategories in Text 2. However, ANOVA results proved that the difference between both groups was only significant in punctuation (p=.0479) and mistranslation (p=.00283) in Text 1, which reveals that the impact of training is not relevant in this pilot study. This result is consistent with Temizöz (2013) according to whom post-editing performance between different educational profiles (engineers and professional translators) were similar in terms of quality.

Research questions number 3 and 4 were related to the impact of directionality on the linguistic correctness and accuracy of post-editing. Specifically, they sought to answer to what extent post-editing into L1 (question 3) and L2 (question 4) performed by groups A and B is linguistically correct and accurate. Comparing the arithmetic means for groups A and B in Texts 1 and 2, both groups obtained more successful edits when post-editing into L1 than when post-editing into L2. Group A (whose L1 is English) obtained more successful edits in the Spanish into English post-editing (AM 46%) than in the English into Spanish post-editing (AM 40%). Group B (whose L1 is

Spanish) obtained more successful edits in the English into Spanish post-editing (AM 74.3%) than in the Spanish into English post-editing (AM 60%). However, as explained above, ANOVA results proved that the difference in successful edits between both groups was only significant in punctuation (p=.0479) and mistranslation (p=.00283) in the English into Spanish post-editing. Therefore, it was not possible to prove that directionality implies differences in quality. This result is consistent with the results in Sánchez-Gijón and Torres-Hostench (2014) and Igor et al. (2017) as both studies did not find significant differences when directionality was involved. However, our result is inconsistent with Garcia (2011) according to whom participants post-editing into L2 worked better (i.e. obtained higher marks) than post-editing into L1.

Research questions 5 and 6 sought to explore the relation between productivity with training (research question 5) and productivity with directionality (research question 6). With regard to training, group B (Translation students) post-edited Texts 1 and 2 in 28.4 minutes whereas group A (Modern Language students) post-edited both texts in 30.4 minutes. Regarding directionality, group B (whose L1 is Spanish) post-edited Text 1 faster (post-editing from English into Spanish) than group A and group A (whose L1 is English) post-edited Text 2 faster (post-editing from Spanish into English) than group B. However, no significant differences were found between the groups and, consequently, the impact of training or directionality on productivity could not be proved.

Due to the limitations of our pilot study, the results cannot confirm whether the fact that both groups were given the same training in post-editing is the key aspect and the explanation behind the similar results. However, this study opens up various lines for further research. The study could be conducted with more participants and with more texts. Another possibility would be to replicate the study with other post-editor profiles and with other language pairs.

References

Aranberri, N., Labaka, G., & Diaz de Ilarraza, A. (2014). “Comparison of Post-editing Productivity Between Professional Translators and Lay Users”. In: O'Brien, S.; Simard, M. & Specia, L. (eds.). Proceedings of the Third Workshop on Post-editing Techniques and Practices (WPTP-3): The 11th Conference of the Association for Machine Translation in the Americas: October 22-26, 2014: Vancouver, BC Canada. [S.l.]: AMTA, pp. 20-33. <https://www.amtaweb.org/AMTA2014Proceedings/AMTA2014Proceedings_PEWorkshop_final.pdf>. [Last accessed on April 17, 2018].

Center for Research and Innovation in Translation and Translation Technology (2018). CRITT Translation Process Database. <https://sites.google.com/site/centretranslationinnovation/tpr-db> [Last accessed on April 17, 2018].

Flanagan, M., & Christensen, T.P. (2014) "Testing post-editing guidelines: how translation trainees interpret them and how to tailor them for translator training purposes". The

Interpreter and Translator Trainer, 8, 2, 257-275. <https://doi.org/10.1080/1750399X.2014.936111> [Last accessed on April 17, 2018].

Garcia, I. (2011). "Translating by post-editing: is it the way forward?". Machine Translation, 25, 217-237. <https://doi.org/10.1007/s10590-011-9115-8> [Last accessed on April 17, 2018].

Igor, L.S., Alves, F., Schmaltz, M., Pagano, A., Wong, D., Chao, L., Leal, A.L., Quaresma, P., Garcia, C., & Da Silva, G.E. (2017). "Translation, Post-Editing and Directionality: A Study of Effort in the Chinese-Portuguese Language Pair". In Jakobsen, A.L., & Mesa-Lao, B. (eds.). Translation in Transition. Between cognition, computing and technology. Amsterdam/Philadelphia: John Benjamins Publishing Company, 91-117.

Kenny, D., & Doherty, S. (2014). "Statistical machine translation in the translation curriculum: overcoming obstacles and empowering translators". The Interpreter and Translator Trainer, 8, 2, 276-294. <https://doi.org/10.1080/1750399X.2014.936112> [Last accessed on April 17, 2018].

Mellinger, C.D. (2017). "Translators and machine translation: knowledge and skills gaps in translator pedagogy". The Interpreter and Translator Trainer, 280-293. <https://doi.org/10.1080/1750399X.2017.1359760> [Last accessed on April 17, 2018].

O'Brien, S. (2002). "Teaching Post-editing: A Proposal for Course Content". In: Sixth EAMT Workshop: Teaching Machine Translation: 14-15 November 2002: UMIST, Manchester: Proceedings, pp. 99–106. <http://mt-archive.info/EAMT-2002-TOC.htm>. [Last accessed on April 17, 2018].

O'Brien, S. (2011). "Towards predicting post-editing productivity". Machine Translation, 25,197-215. <https://doi.org/10.1007/s10590-011-9096-7> [Last accessed on April 17, 2018].

Plitt, M. & Masselot. F. (2010). "A Productivity Test of Statistical Machine Translation Post-Editing in a Typical Localisation Context". The Prague Bulletin Mathematical Linguistics, v. 93, n. 1, pp. 7-16. <https://doi.org/10.2478/v10108-010-0010-x>. [Last accessed on April 17, 2018].

Sánchez-Gijón, P. & Torres-Hostench, O. (2014). "MT Post-editing into the Mother Tongue or into a Foreign Language? Spanish-to-English MT translation output post-edited by translation trainees". In: O'Brien, S.; Simard, M. & Specia, L. (eds.). Proceedings of the Third Workshop on Post-editing Techniques and Practices (WPTP-3): The 11th Conference of the Association for Machine Translation in the Americas: October 22-26, 2014: Vancouver, BC Canada. [S.l.]: AMTA, pp. 15-19. <https://www.amtaweb.org/AMTA2014Proceedings/AMTA2014Proceedings_PEWorkshop_final.pdf>. [Last accessed on April 17, 2018].

Snover, M., Dorr, B., Schwartz, R., Micciulla, L. & Makhoul, J. (2006). "A study of translation edit rate with targeted human annotation". In Proceedings of the Fifth Conference of the Association for Machine Translation in the Americas (AMTA), pages 223–231, Boston, MA.

TAUS. (2013). TAUS Best Practice Guidelines. Quality Evaluation using an Error Typology Approach. <https://www.taus.net/academy/best-practices/evaluate-best-practices/error-typology-guidelines>. [Last accessed on April 17, 2018].

Temizöz, O. (2013). Post-editing Machine Translation Output and its Revision: Subject-Matter Experts versus Professional Translators [Doctoral thesis]. Universitat Rovira i Virgili, Tarragona. <http://www.tdx.cat/bitstream/handle/10803/128204/TemizozOzlem_TDX2.pdf?sequence=1>. [Last accessed on April 17, 2018].

Ulrich, G. (2013). Yet Another Word Alignment Tool (YAWAT). <http://dighum1.ftsk.uni-mainz.de/cgi-bin/yawat/yawat.cgi> [Last accessed on April 17, 2018].

Yamada, M. (2014). "Can college students be post-editors? An investigation into employing language learners in machine translation plus post-editing settings". Machine Translation, 29(1), 49-67. <https://doi.org/10.1007/s10590-014-9167-7> [Last accessed on April 17, 2018].