Language Learning & Technology
ISSN 1094-3501 CC BY-NC-ND

2022, Volume 26, Issue 1
pp. 1–17

ARTICLE

Training in machine translation post-editing for foreign language students

Hong Zhang, Yangzhou University

Olga Torres-Hostench, Autonomous University of Barcelona

Abstract

The main purpose of this study is to evaluate the effectiveness of Machine Translation Post-Editing (MTPE) training for FL students. Our hypothesis was that with specific MTPE training, students will able to detect and correct machine translation mistakes in their FL. Training materials were developed to detect six typical mistakes from Machine Translation (MT) raw output: Accuracy, Word Order, Official Name, Preposition, Omission, and Formal Style. The training materials include three levels of difficulty: Initial - ability to spot a mistake, Intermediate - ability to classify the type of mistake, and Advanced - ability to correct the mistake. A pretest-posttest design with a control group and a trained experimental group was chosen to test the effectiveness of the training programme. In the posttest, the experimental group could identify and correct more mistakes successfully. *and in less time than the control group, especially for* omission, official *name and* preposition. Accuracy, formal style, *and* word order *errors were more difficult to correct. Results suggest that specific MTPE training is not only useful to identify and correct MT mistakes but also a way to incorporate a critical view on machine translation in FL classes.*

Keywords: *Computer-Assisted Language Learning, Machine Translation, Foreign Language Learning, Second Language Acquisition*

Language(s) Learned in This Study: *Chinese, Spanish*

APA Citation: Zhang, H., & Torres-Hostench, O. (2022). Training in machine translation post-editing for foreign language students. *Language Learning & Technology*, *26*(1), 1–17. http://hdl.handle.net/10125/73466

Introduction

The use of Machine Translation (MT) as a resource for foreign language production is becoming popular among students, but less so among teachers. Nevertheless, when language students put their acquired foreign language skills into practice in specialized contexts such as international trade or global business, they may well use MT when writing to international customers or translating a price list, for instance. In our view, instead of just banning MT in academic work, teachers could use it as an additional teaching resource in the classroom, and specific training may help students to use it critically.

While this article has a general focus on machine translation post-editing (MTPE) training for foreign language (FL) students, it looks specifically at assessing the usefulness of post-editing (PE) training materials produced for Chinese L1 students studying Spanish as their L2 (B2 level), based on the hypothesis that MTPE training could help students identify and correct raw MT output. We address three research questions:

1. To what extent can MTPE training help students identify and correct raw MT output? [This was tested with a pretest-posttest evaluation]
2. What kinds of MT mistakes are easiest to correct?
   [This was tested by observing which MT errors are identified and corrected by more students]
3. What kinds of MT mistakes are most difficult to correct?
   [This was tested by observing which MT errors are identified and corrected by less students]

Literature Review

The study of MT in language learning has been investigated by different authors and from different perspectives for decades. As far back as 1995, Anderson argued that “MT can be used as a powerful focal point in L2 learning” (Anderson, 1995, p. 89). Moving forward in a chronological overview of the literature, another noteworthy study is that of Kliffer (2005), which found not only that students post-editing into L2 improved MT output remarkably, but also that the weaker students valued the experience most and in fact preferred PE to translating from scratch. Perhaps the most cited article on this issue is by Niño (2008), who stated that through target language mistake detection and correction, MTPE proved beneficial for advanced students. Later, Niño also suggested that using MT for advanced language students could help their “awareness as to the complexity of translation and language learning” (Niño, 2009, p. 253). Another study in which MT proved helpful for language learning was that of Clifford et al. (2013, p. 116), where students critically assessed MT output, recognizing that, while it contained mistakes, they found MT to be helpful in their language learning, especially for vocabulary acquisition.

In relation to analysing MT mistakes and using PE in language learning, Kliffer stated that the post-editing of MT “gave students insight into the huge challenges which have confronted MT, especially the questions of how to deal with syntactic and lexical ambiguity, non-literal language, and inferencing” (Kliffer, 2008, p. 63). Another interesting study on MT mistakes was conducted by Fredholm (2015), who looked at the mistakes pupils made when using online translation; they made fewer spelling or article/noun/adjective agreement errors, but more syntax and verb morphology mistakes. The studies in question point to specific training being necessary, as suggested by Sycz-Opoń and Galuskina (2017), who stated that post-editing raw MT output requires critical thinking and perceptiveness, and that training in the use of MT technology should be implemented in translation classes. In line with this approach, Rico et al. (2017) presented a PE training proposal in which students not only learn basic PE techniques but also question their preconceptions of MT to some degree.

Notwithstanding, there are also studies that discourage the use of MT in language learning, such as those of Loffler-Laurian (1983, 1985) and Lewis (1997). García and Pena (2011) noted that using MT does not help students because it could make them more dependent on technology than knowledge. Also, Fredholm (2015) advised that while advanced students can use MT, it is counterproductive for beginners and intermediate students. Additionally, there is an ongoing debate on the suitability of PE training for L1 speakers and FL students. Sánchez-Gijón and Torres-Hostench (2014) compared the PE skills of English L1 and English L2 translation trainees and the results were promising. L2 students were able to identify omission and mistranslation mistakes even better than English L1 translation trainees, but found it more difficult to identify grammar and syntax mistakes. In a nutshell, MTPE’s potential for FL training is a field that is well worth exploring to obtain more data and information.

In order to design the present study for L2 students, it was necessary to consider not only previous research, but also the language combination involved and the quality of existing raw MT output for that language combination. Preliminary preparation of the study included two tasks. The first was a comparison of three MT engines, Google Translate, Baidu, and Bing Translator, to determine which was best suited to the language combination Chinese > Spanish. Google Translate produced the highest quality results (Zhang, 2016) and was thus the chosen engine for our study design. The second preliminary task consisted of a pilot test involving work on raw MT output containing numerous MT mistakes of different kinds, the aim being to identify which of them were easiest or most difficult to correct for Chinese L1 students of Spanish as their L2. The participants found detecting mistakes very difficult (Zhang, 2017) and we learnt that it would not be advisable to expose students to MT without previous MTPE training, as well as that it might be more useful to focus on some specific error types. We found that most of the participants did not know how to identify or correct the errors in the raw MT output. We therefore drew on the pilot test to identify a series of relevant MT mistakes on the basis of which training materials could be developed. In other language combinations, the chosen MT mistakes may need to be different. It must be said that Chinese and Spanish are distant languages, and MT systems often use English internally as an intermediate language (i.e., Chinese is internally translated into English and English into Spanish), so MT results may be disappointing.

Method

The present study was designed taking into consideration the above-mentioned literature and preliminary tasks. We chose a pretest-posttest design involving taking measurements both before and after a training programme in MTPE. There were two different groups of participants: a control group (who were not given any training) and an experimental group (who took a training programme in MTPE). The study can be considered quasi-experimental as the participants were not assigned randomly to each group. All participants took the pretest at the same time, after which volunteers were asked to attend the training sessions. These volunteers became the experimental group. This means that the experimental group would have been more motivated to learn, and this is factored into the results. Overall, sixteen students participated: eight in the control group and eight in the experimental group. Their profile is highly homogeneous: there were four males and twelve females, all of them from China and with a degree in Spanish Studies from China and a B2 level of Spanish as L2 according to the Common European Framework of Reference (a B2 certificate being required to access the master's degree course in translation they were all currently taking in Spain). They were aged between 23 and 25. None of them had previously taken MTPE training. Before the pretest, they gave their informed consent to participate in the study. Upon finishing the pretest, they filled in a brief demographic information questionnaire. The total time allowed for the pretest and postest was 90 minutes (pretest = 45 minutes, posttest = 45 minutes). BB FlashBack screen recording software was used to make a note of every action taken by the students. The participants could use the Internet as an aid and all their searches were recorded by BB FlashBack.

The instrument used for the pretest and the posttest was the same. It consisted of ten L1 (Chinese) sentences translated into L2 (Spanish) by Google Translate. Nine of them each contained one mistake while the tenth contained no mistakes (see Appendix). Each incorrect sentence contained only one error for didactic purposes. The error categories were as follows: (a) accuracy, (b) word order, (c) official name, (d) preposition, (e) omission and (f) formal style (described in more detail later). The MT mistakes were chosen from the aforementioned pilot study, but it should be noted that between the pilot test and the tests performed for the present study Google Translate's algorithms changed and so too did the errors it produced. From 2006 to 2017, Google used statistical MT engines that produced different mistakes than the neural MT engines it currently uses, which provide better output. This study was carried out with the current Google neural translation engine, so the chosen MT mistakes are relevant to neural MT as well as to the students' level. The pretest trial was performed by two students (neither of them participants in the present study) and one lecturer in order to verify suitability and appropriateness to the students' language level.

In order to analyse the results of the pretest and posttest tasks, the following scores were established:

- Student failed to identify the MT mistake (0).
- Student highlighted the mistake but did not edit it (1).
- Student highlighted the mistake but did not correct it successfully (2).
- Student identified and edited the mistake correctly (3).

The results obtained allowed us to analyse successful edits versus required edits, and the number of MT mistakes that students identified and corrected successfully. After the pretest, the experimental group took specific training on MTPE. The idea of specific training was chosen after reading the work of authors such as Sycz-Opoń and Galuskina (2017) and Rico et al. (2017), who suggested this kind of approach. A few days after the training, and a month after the pretest, the posttest was performed by both groups together.

MTPE Training Session Design

Chinese L1 students are used to learning foreign languages by doing a lot of grammar exercises, so the idea of MTPE training for Chinese students was based on this tradition and exercises were prepared with examples of types of mistakes and how to solve them. Moreover, in China, there is an exam named "Test of Professional Spanish – Level 4" which contains a section in which students have to identify and correct the mistakes in sentences in Spanish. So, the training sessions took this into consideration, and it was developed around identifying and correcting specific mistakes.

The training proposal has two noteworthy features, which can be transferred to any MTPE training course in any language combination. The first is the list of specific types of mistakes. The training did not consist of correcting *any kind of* MT mistake, but rather previously identified MT mistakes that, with due training, students can spot and correct. This *controlled* training develops confidence in the students as they focus in detecting specific mistakes. This list of mistakes is as follows:

- ***Accuracy***, defined in MQM (Lommel et al., 2014) as "The target text does not accurately reflect the source text;" this refers to target text lexical accuracy. Participants were asked to spot differences in meaning between the source text and the MT.
- ***Official name***, namely proper names, names of entities, places, and so forth that MT translates incorrectly and are mostly terms that are more encyclopaedic or cultural than common names. The participants were trained to check all proper names and not trust MT output.
- ***Preposition,*** specifically misuse of prepositions and so on. This was chosen as an example of a grammar issue that can be learned by the trainees.
- ***Word order***. This is a quite common MT error from Chinese into Spanish because word order in both languages is quite different. Participants were instructed to focus on this issue.
- ***Formal style***, defined in MQM (Lommel et al., 2014) as "Register". For example, "The text uses a level of formality higher or lower than required by the specifications or general language conventions." MT cannot keep a homogeneous style through a text, but with training, participants were able to spot register incoherence in the text. We renamed this mistake from "Register" to "Formal style" because we wanted the students to think about the formality of the text.
- ***Omission***, defined in MQM (Lommel et al., 2014) as "Content is missing from the translation that is present in the source." Participants were trained to spot omissions in the MT output.

The second feature is the level of difficulty classification:

1. Level 1 (Initial). Mistake detection: The student can detect a MT mistake.
2. Level 2 (Intermediate). Mistake typology: The student can classify the type of mistake. This involves a greater linguistic awareness.
3. Level 3 (Advanced). Mistake correction: he student can successfully correct the identified mistake, which would be the level for advanced students.

The training proposal includes examples for the six types of mistakes (accuracy, word order, official name, preposition, omission, formal style) as well as for the three levels of difficulty, so that students can improve their MTPE skills step by step.

The source sentences for the examples are taken from the book by Sheng (2006), a book recommended by the Chinese government for teaching Spanish. This book is a translation course (Spanish-Chinese) built on a global and integrated conception of advanced students' learning whose objective is to foster both their communicative competence and their personal development. Multiple Chinese sentences in this book were Google-translated into Spanish by the authors of the study, and those, which clearly illustrated the aforementioned relevant types of mistakes, were chosen for the pretest-posttest and for the training sessions.

The main content of this voluntary 10-hour training course consisted of two 2-hour sessions in class and six hours of self-learning (four hours to identify and correct mistakes, one hour for a specific exercise on Spanish prepositions (based on detection of needs from the prettest), and one hour to check the solutions). The time given over to this course was limited by pragmatic considerations at our university and a longer training period would have been preferable. However, the 10 hours were deemed sufficient to have an impact on the skills of the students. The contents of the MTPE training course were as follows:

- Session 1. Introduction to the six types of mistakes in MT from Chinese into Spanish. The first session was mainly practical in order to engage the participants and so they could appreciate the usefulness of the course. The six types of mistakes were explained with an example, and then students were invited to work in pairs and detect and correct other examples of the same type. The three levels of difficulty criteria explained above were presented at this first session.
- Session 2. Introduction to theoretical content on MTPE. Now that the students had been introduced to the possibilities of MTPE training after the error detection exercises, the students were given some theoretical background on what MT is, types of MT, what PE is, what a post-editor does, levels of PE, norms and guidelines for PE, PE as a professional career for someone with foreign language skills, and relevant bibliography.
- Self-learning. Participants received an activity dossier with four lessons by e-mail with the solutions and they practiced the exercises on their own.

Training materials have been uploaded to the IRIS Database (Mackey and Marsden, 2015), the digital repository of instruments and materials for research into second languages. In specific terms, the following can be found this database:

- PowerPoint file for a seminar on MTPE training for FL students with the following content: list of specific MT mistakes; examples for each error type (omission, word order, official names, accuracy, formal style and preposition); examples in Chinese and neutral Spanish (suitable for different Spanish varieties); examples according to the three levels of difficulty: (a) identifying the mistake (first level), (b) classifying the mistake (second level), and (c) correcting the mistake (third level).
- A PowerPoint presentation for FL students on the basics of MTPE including: (a) MT definition, (b) types of MT, (c) definition of PE, (d) usefulness of PE, (e) levels of PE, (f) PE tasks and PE recommendations, (g) pros and cons of PE, and (h) bibliographical references.
- A student dossier with four sessions of self-learning exercises and solutions to post-edit MT output from Chinese into Spanish.

After the training course, participants were asked for their opinion and most participants valued it very positively. Negative opinions were related to the short length of the training (only 10 hours) as they considered more time was necessary to become familiar with the error types and the examples. Participants recognized that the error types for MT were also a problem in their own L2 written expression and they would like more practice to correct or avoid these errors. They also acknowledged they quite often use MT and being able to identify and correct its mistakes would be very useful for them. A few days after the course, control and experimental groups were invited to do the posttest, the results of which are presented later.

Results and Discussion

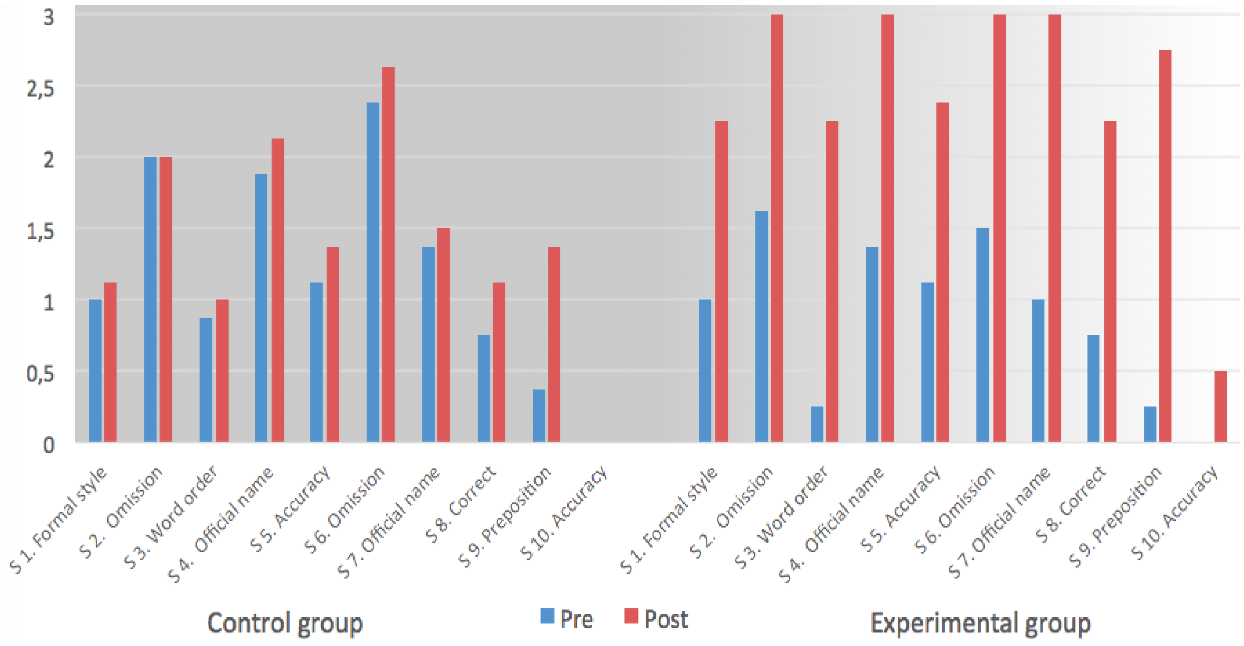
This section includes results and discussion of the identification and correction of MT mistakes, the number of successful edits versus the required edits, results on the lengths for pauses (see further context on pauses under the results section on pauses), the types of pauses, and the number of pauses. Several forms of quantitative analysis were carried out. Descriptive statistics were used, specifically frequencies for qualitative variables and mean and standard deviation for quantitative variables. Comparisons between pretest and posttest were performed for continuous variables, using student's t-test when conditions (normality and homoscedasticity) were satisfied and the Wilcoxon signed-rank test when they were not. The statistical analysis was performed using SPSS software, using a nominal significance level of 5% ($p < 0.05$).

Scores for Identifying and Correcting MT Mistakes

Figure 1 encapsulates the contribution of this paper to language learning studies and is the final result of the study. It summarizes the results for identifying and correcting mistakes between the pretest (blue columns) and the posttest (red columns), and between the control group (on the left) and the experimental group (on the right). Values ranged from 0 (mistake not identified) to 3 (mistake identified and edited correctly), with 1 being (mistake identified but not edited) and 2 (mistake identified but edited incorrectly). Figure 1 shows that the control group scored higher in the pretest than the experimental group, but their posttest results are close to pretest scores (except for the preposition mistake). In contrast, the experimental group scored lower than the control group in the pretest, but after the training course, their scores are much higher than those of the control group. It is clear to see that the students who received specific training on MT mistakes (experimental group) achieved better results than those in the control group.

Figure 1

Scores for Correcting MT Mistakes (Pretest vs. Posttest / Control vs. Experimental)



Focusing on the different MT mistake types, in Table 1 there is a detailed comparison of the mean values of the score between pretest / posttest for the control group vs. experimental group by sentence. Possible values ranged from 0 (mistake not identified) to 3 (mistake identified and edited correctly). As seen in Table 1, not all the participants were able to identify and correct the six types of MT mistakes from the ten sentences. A comparison of the pretest / posttest mean value scores shows that the experimental group scored higher in the posttest. They found accuracy, formal style, word order, and preposition mistakes difficult to detect in the pretest, while after the MTPE training, the experimental group did well at identifying and correcting word order, omission, preposition and official name. However, accuracy and formal style proved difficult for both groups. Despite the mistakes being specific to the analysed sentences, these results clearly show a pattern of better performance among the students who received specific training. The study of MT mistakes detected by language students has also been analysed by Fredholm (2015), Clifford et al. (2013), Belam (2003) and Kliffer (2008).

Table 1

Mean Scores for Correcting MT Mistakes (Pretest vs Posttest / Control Group vs. Experimental Group)

Mistake type in each sentence	Mean (control group)			Mean (experimental group)			*p* - Value[b]	*p* - Value[c]
	Pre	Post	*p* - Value[a]	Pre	Post	*p* - Value[a]		
S1. Formal style	1.00	1.13	0.351	1.00	2.25	0.0001*	-	1.000
S2. Omission	2.00	2.00	1.000	1.63	3.00	0.0001*	0.074	0.167
S3. Word order	0.88	1.00	0.351	0.25	2.25	<0.0001*	0.071	0.043*
S4. Official name	1.88	2.13	0.171	1.38	3.00	0.0015*	0.421	0.156
S5. Accuracy	1.13	1.38	0.171	1.13	2.38	0.0190*	0.500	0.500
S6. Omission	2.37	2.63	0.171	1.50	3.00	<0.0001*	0.201	0.009*
S7. Official name	1.36	1.50	0.351	1.00	3.00	0.0005*	0.289	0.193
S8. Correct (no error)	0.75	1.13	0.351	0.75	2.25	0.0331*	0.500	0.500
S9. Preposition	0.38	1.38	0.001*	0.25	2.75	<0.0001*	0.388	0.309
S10. Accuracy	0.00	0.00	1.000	0.00	0.50	0.1710	-	1.000

Note. *p < 0.05, P values below 0.05 are marked with an asterisk / a: Wilcoxon signed-rank test / b: F test / c: ANOVA test of two groups in the pretest.

Interestingly, in the control group there was a significant statistical difference between pretest and posttest only in one sentence with a preposition mistake (p = 0.001, p < 0.05). As for the experimental group, there were significant statistical differences between pretest and posttest for nine out of ten sentences. This suggests that MTPE training helped the experimental group to identify and correct the mistakes of all different error types (official name, preposition, word order, formal style and omission). Some specific results deserve closer attention:

- Omission mistakes: the highest scores were recorded for identifying and correcting these mistakes in both pretest and posttest for both groups. It seems clear that students are good at spotting omissions intuitively (addressing the research question posed concerning which mistakes students were able to detect with the most ease).
- Word order: scores are low in the pretest for both groups. It appears that students are unwilling to change MT output word order. Nevertheless, the scores from the experimental group in the posttest are very high, which means that with proper training, they overcome this reluctance.
- Official name: the experimental group scored lowest in the pretest yet highest values in the posttest. The training sessions raised their awareness of this particular MT limitation and the need to always check official names.
- Accuracy: these results are low for both groups. As regards sentence 10, after analysing the results the authors realized that the example was not appropriate, so low results in both groups and in pretest-posttest are probably not due to students' lack of skills. A better sentence for replication purposes has been included in the Appendix. Nevertheless, accuracy is such an open

type of error that it is more difficult to spot as it is related to sentence meaning rather than to specific misuse of words. More and better training materials must be developed to help students identify and correct accuracy errors.

- Preposition: results for the experimental group are really outstanding from 0 (not detected) to 3 (correctly identified and edited).
- Correct sentence: the control group added more errors in the posttest, while the experimental group achieved good results and did not add errors to the correct sentence. For example, one student used an incorrect verb form in the sentence "La ceremonia (…) llevará a cabo a las 7 pm (…)". The correct sentence would be "La ceremonia (…) se llevará a cabo a las 7 pm (…)" [Ceremony (…) will take place at 7 pm (…)].

As for the complementary statistical information on Table 1, in the ANOVA test there were statistical differences between the control group and the experimental group in two pretest sentences: S3 (word order) and S6 (omission) ($p = 0.043$ and $p < 0.009$). This means that before the MTPE training session, the control group did much better in terms of word order (0.88 vs. 0.25) and omission (2.38 vs. 1.50) mistakes than the experimental group did in the pretest. Additionally, the control group's pretest values can be seen to be always equal to or better than the experimental group's values.

There are at least two possible reasons for the experimental group correcting more mistakes in the posttest than the control group. First, the MTPE training was useful and helped them to identify and correct mistakes. Second, the experimental group was comprised of volunteers, and their curiosity meant they made the most of the training given on MTPE. Interestingly, although the experimental group scored lower than the control group in the pretest, their posttest results were much better than those of the control group.

Results - Edits

Lacruz et al. (2014) report that the number of required edits refers to the least number of insertions, deletions, substitutions, and shifts required to convert the MT output into the final post-edited version. Table 2 shows the number of all edits by the participants vs. the required (successful) edits. It should be pointed out that the total required number of edits was nine, since one sentence contained no mistakes and there was only one mistake in each of the other sentences. Any result above nine means that participants made unnecessary changes.

Table 2

Number of Total Edits vs. Required Edits (Pretest vs Posttest / Control Group vs. Experimental Group)

	Control mean			**Experimental mean**				
	Pre	Post	*p* – Value[a]	Pre	Pos	*p* – Value[a]	*p* – Value[b]	*p* – Value[c]
Mean for participants' total edits	11.50	16.00	0.018*	13.50	14.13	0.493	0.321	0.149
Mean for participants' successful edits	4.38	4.88	0.726	3.00	7.00	0.011*	0.733	0.081

Note. *$p < 0.05$, p values less than 0.05 are marked with one asterisk / a: Wilcoxon signed-rank test / b: Levene test / c: ANOVA test

As shown in Table 2, the Levene test revealed a statistical difference between the control group and the experimental group (p = 0.149, and p = 0.081). In Table 2 above, the Wilcoxon signed-rank test was used to compare the control group and the experimental group. In the experimental group, there was a significant statistical difference between pretest and posttest (p = 0.018) for the number of post-edits, which shows that the control group performed more edits in the posttest. One reason may be that the control group already knew the sentences from the pretest and tried to fix them with more edits. However, notice that their edits were mostly unsuccessful.

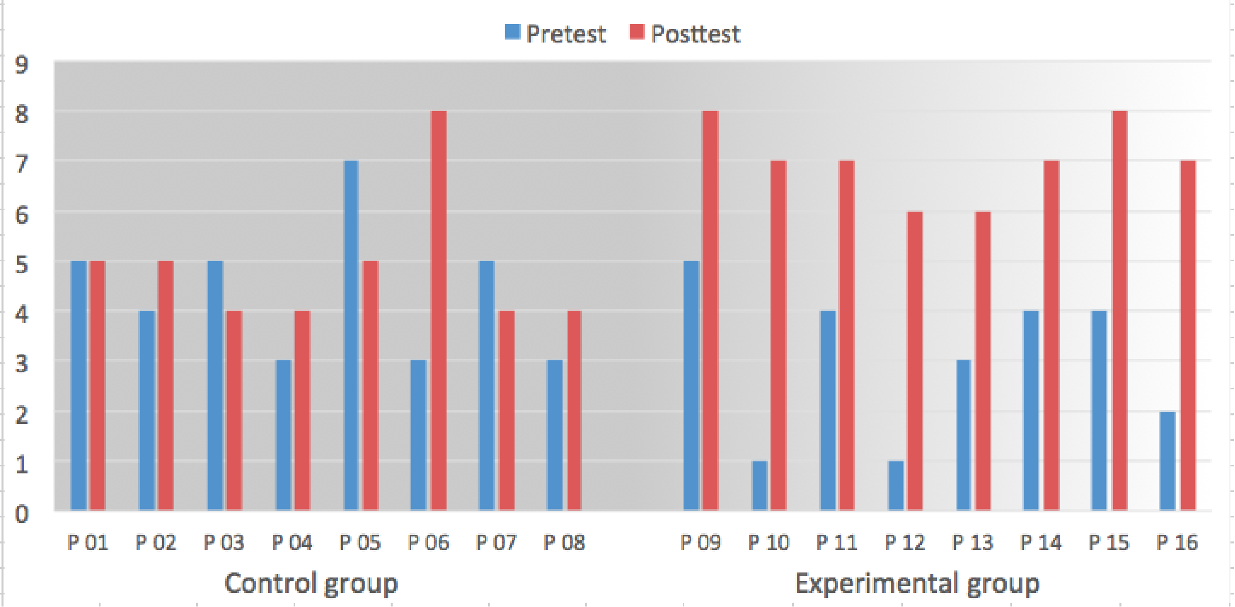
Regarding the experimental group, there was no significant statistical difference between pretest and posttest (p = 0.493) for the total number of edits. But there was a statistical difference in the number of successful edits between pretest and posttest in the experimental group (p = 0.011), which means that they took advantage of the training sessions. In the comparison between the mean value of the number of required edits of the two groups, the experimental group performed twice as many successful edits compared to the pretest (seven vs. three). Actually, the required number of edits was nine, so seven is a remarkable result.

In the experimental group, there were significant statistical differences between pretest and posttest successful edits for formal style, word order, accuracy, and preposition (i.e., in five out of ten sentences).

As for the individual performance of participants, Figure 2 shows the total number of successful edits by each participant (vertical axis). This also clearly shows how the control group performed better in the pretest while the participants of the experimental group improved in the posttest.

Figure 2

Total Number of Successful Edits (By Each Participant).



García and Pena (2011) remind us that successful edits can improve the target text, but unsuccessful edits may also add mistakes. In the screen recordings taken from the pretest and posttest for each participant, it was noticed that in the pretest, most of our participants searched the Internet for prepositions, grammar, and other linguistic structures, but in the end they did not modify the original sentence because of their doubts about the foreign language, and hence the edit was unsuccessful. In the posttest, trained students knew what to do and what to look for, so they were more successful.

Results - Total Time

For the control group, the mean values for total time of pretest (2061 s) were higher than those of the posttest (1571 s), and for the experimental group, the mean values of total time of pretest (2327 s) were higher than those of posttest (1411 s). If we look at the maximum and the minimum values for each participant, it can be observed that the time differences among participants were quite high. In the experimental group, the times ranged from 1796 to 2703 seconds in the pretest and 1109 to 1751 seconds in the posttest. In the control group, the times ranged from 1540 to 2399 seconds in the pretest and from 874 to 2056 seconds in the posttest.

Results – Pauses

Under the experimental conditions, the participants were allowed to consult the Internet for information during the PE session. However, typing pauses attributable to these searches are different in nature to typing pauses that take place in professional PE settings, where professionals use pauses mainly to think. Pauses, measured by keystroke logging or by eye tracking data on fixations and gaze duration, are known to be good indicators of cognitive demand in monolingual language production (Schilperoord, 1996). The pause times in this study are extremely high. With data obtained via keylogging and screen recordings, it cannot be stated that the participants were making a cognitive effort during the entire duration of the pause (because we know they were searching the Internet, for instance). However, it can be stated that they were extremely insecure about PE and might not have had enough L2 knowledge to correct the MT mistakes.

As a preliminary observation, it can be said that PE mistakes were not due to hasty decisions but to lack of knowledge. Pauses were observed by means of BB FlashBack recordings (as used by García & Pena, 2011). In Table 3, the Wilcoxon signed-rank test was used to compare the control group with the experimental group.

Table 3

Control Group and Experimental Group Pause Time by Sentence in Pretest and Posttest

Mistake Type in Each Sentence	**Control Group (in Seconds)**					**Experimental Group (in Seconds)**				
	Pre		**Post**		***p* – Value**^{a}	**Pre**		**Post**		***p* – Value**^{a}
	M	SD	M	SD		M	SD	M	SD	
S1. Formal style	277	212	122	63	0.069	224	148	81	77	0.025*
S2. Omission	177	80	140	72	0.401	144	103	110	67	0.401
S3. Word order	319	200	136	62	0.161	407	261	137	101	0.025*
S4. Official name	27	14	45	12	0.263	31	12	20	31	0.107
S5. Accuracy	224	71	222	19	0.779	179	94	90	149	0.050*
S6. Omission	111	142	119	119	1.000	246	44	160	74	0.161
S7. Official name	73	71	117	116	0.183	111	57	151	48	0.779
S8. Correct	167	124	151	90	0.889	196	90	157	111	0.161
S9. Preposition	209	148	183	37	0.575	244	76	90	142	0.017*
S10. Accuracy	162	53	240	113	0.483	98	102	170	106	0.021*

Note. *$p < 0.05$, p values less than 0.05 are marked with one asterisk / a: Wilcoxon signed-rank test / M: mean / SD: standard deviation

In the control group, there were no significant statistical differences between the pretest and the posttest pause times, which means that the group found the difficulty of the task similar on both occasions. Turning to pause times in each sentence, the participants spent a lot of time on every pause, which suggests that PE was difficult for them. Particularly in the case of official name (S7) and accuracy (S10), the posttest time mean value was higher than that of the pretest, which means these sentences were more difficult for the students. Regarding deviation, for the experimental group, in general, the value of the standard deviation mean was lower in the posttest (94 seconds) than in the pretest (137 seconds); in contrast the standard deviation mean between the pretest and posttest for these ten sentences was more dispersed in the control group (123 seconds vs. 113 seconds). This might mean that in the experimental group there is more coherence between the detection time and the correction after training. Additionally, these results show that with MTPE training, the experimental group spent less time on PE mistakes related to formal style, word order, accuracy and preposition.

Furthermore, a greater number of pauses did not mean a greater modification of sentences, and was even less indicative of correct editing. It was noticed that after the MTPE training, participants edited more successfully during pauses.

Besides the pause time, it is interesting to look at the total number of pauses. Table 4 shows the relationship between the number of pauses and the kind of mistake the participant was looking at. This has been analysed by observing screen recordings. In this study, a pause longer than 3-seconds was considered a "pause." In Table 4, the Wilcoxon signed-ranks test was used to compare six types of mistakes between the control group and the experimental group. In the control group, there was a significant statistical difference between pretest and posttest concerning word order (*p* = 0.028), omission (*p* = 0.041), and formal style (*p* = 0.018), which means the group made fewer pauses in the posttest. The ratio was 1:2. In contrast, there were no significant statistical differences between pretest and posttest in the case of the experimental group for any of these six types of mistakes. As shown also in Table 4, we found that in the HOV test, there was a significant difference between the control group and the experimental group for the formal style mistake, so the ANOVA test was performed as shown in Table 5. The number of pauses corresponding to the formal style sentence differed considerably in the pretest and less so in the posttest.

Table 4

Mean Number of Pauses by Mistake Type (Pretest vs. Posttest / Control Group vs. Experimental Group)

	Control (pauses mean)			**Experimental (pauses mean)**			***P* - Value[b]**	***p* - Value[c]**
	Pre[d]	**Post[d]**	***p* - Value[a]**	**Pre[d]**	**Post[d]**	***p* - Value[a]**		
Accuracy	12.13	10.75	0.230	9.88	8.25	0.362	0.392	0.323
Word order	7.13	2.88	0.028*	4.63	2.50	0.156	0.783	0.255
Official name	3.00	2.00	0.290	4.25	2.63	0.481	0.258	0.455
Preposition	4.25	2.13	0.090	3.25	1.63	0.062	0.438	0.446
Omission	5.13	2.25	0.041*	3.25	1.88	0.138	0.531	0.175
Formal Style	6.38	1.25	0.018*	3.75	1.75	0.063	0.250	0.049*

Note. *p < 0.05, p values less than 0.05 is marked with one asterisk / a: Wilcoxon signed-rank test / b: Homogeneity of variance test (HOV test) / c: ANOVA test / d: number of PE pauses for each mistake

Table 5

ANOVA Test of Two Factors Between Group Variance and Training Variance

	p - Value[a]	*p* - Value[b]	Adjust R^2
Formal Style	0.137	<0.001	0.460

Note. a: Group variance / b: Training variance

It can be stated that the highest number of pauses in both groups was found in sentences with accuracy problems in the pretest as well as the posttest (it must be also taken into account the fact that there were two sentences with accuracy problems). The results of the study indicate that the MTPE training did not affect the number of pauses of both groups in the post-test.

From these results, the authors plan to prepare more teaching materials to help L1 Chinese students learn Spanish by identifying and correcting errors in MT output. In conclusion, the results of the study answer the initial research questions, as summarized below:

1. *To what extent can MTPE training help students identify and correct raw MT output?*
   Results suggest that specific training in MTPE can improve FL students' ability to identify and correct raw MT output.
2. *What kinds of MT mistakes are easiest to correct?*
   In the pretest, omission mistakes were easiest to correct. In the posttest, omission, official name, and preposition were easiest for the experimental group. As for the control group, prepositions were easiest to correct.
3. *What kinds of MT mistakes are most difficult to correct?*
   In the pretest, accuracy, word order and preposition were the most difficult. In the posttest, accuracy, formal style and word order were the most difficult for both groups.

Taking into account the results of this study, the authors have a number of recommendations for language teachers:

1. PE training may be an effective way of helping identify and correct MT errors.
2. Designing and testing specific activities for MTPE in language classes is better than asking students to correct a random MT text containing all kinds of mistakes.
3. Reflection on MT mistakes is necessary to avoid uncritical use of MT.
4. Comparing MT systems is useful for generating awareness of different MT quality levels and different MT mistakes.
5. Share all teaching materials developed so that other teachers may use them. For instance, the teaching materials used in this study may be downloaded from the IRIS repository of instruments and materials for research into second languages.
6. Lastly, keep abreast of MT developments. MT is evolving quickly and teachers have to be aware of its successes and pitfalls.

Conclusion

This article has focused on evaluating MTPE training, designed specifically for Chinese L1 students of Spanish L2, to help them identify and correct raw MT output. The research design consisted of a pretest-posttest design composed of ten raw MT output sentences with six different types of MT mistakes: accuracy, word order, official name, prepositions, omission and formal style.

The results indicated that PE of MT into L2 turned out to be a very difficult task, but with appropriate training, students can improve their PE skills. In the pretest, most participants could neither identify nor correct the MT mistakes in regards to accuracy, word order, official name, preposition, omission, and formal style. They edited more than was required, even words that were correct. The pretest showed that language students did not know what to look for in the sentences and felt insecure about their language skills. In the posttest, the control group performance was similar to that of the pretest. In contrast, the trained experimental group performance showed statistically significant improvements in nine out of ten sentences in the posttest. Moreover, they spent less time in the posttest, they made fewer pauses, and their edits were more effective. Despite the satisfactory results of the study, some limitations need to be pointed out: the limited number of participants (eight for each group) and the limited time dedicated to training (10-hours of classes and self-study combined). It is likely that the results would have been even better with more time and class sessions.

As for the training proposal, the study demonstrated that training for specific MT mistakes and considering different levels of difficulty may be a suitable approach to PE training. If students are trained in specific error types, they can gain practice and experience in spotting and solving errors. Development of MTPE skills requires focused attention and advanced reading comprehension skills, considered valuable skills for language learning. PE activities might help lecturers gain a greater understanding of students' problems and be useful for FL learning. Results show that repeated practice detecting and correcting MT focused on specific error types is successful. FL students are used to detecting grammar errors in their language exercises, so language teachers could apply this methodology to the detection of MT mistakes. Specific training in MTPE may also be a way go beyond demonizing MT and turn it into a didactic tool useful for practicing specific areas of a language, as well as a method to raise students' awareness of the perils of using MT without criteria. Future studies should investigate training for different language combinations, different levels of language competence, and different types of MT mistakes.

Acknowledgements

This work was partly supported by the DESPITE-MT project "Description of Posteditese in Machine Translation," grant number PID2019-108650RB-I00 [MICINN], and the Jiangsu Provincial Double-Innovation Doctor Program (JSSCBS20211063).

References

Anderson, D. D. (1995). Machine translation as a tool in second language learning. *CALICO Journal*, *13*(1), 68–97. https://eric.ed.gov/?id=EJ522959

Belam, J. (2003). Buying up to falling down: A deductive approach to teaching post-editing. In *Proceedings of workshop on teaching translation technologies and tools, Ninth Machine Translation Summit*, (pp. 1–10). New Orleans, USA. https://aclanthology.org/volumes/2003.mtsummit-tttt/

Clifford, J., Merschel, L., & Munné, J. (2013). Surveying the landscape: What is the role of machine translation in language learning? *@Tic. Revista D'innovació Educativa*, *10*, 108–121. https://doi.org/10.7203/attic.10.2228

Fredholm, K. (2015). Online translation use in Spanish as a foreign language essay writing: Effects on fluency, complexity and accuracy. *Revista Nebrija de Lingüística Aplicada a la Enseñanza de las Lenguas*, *18*, 7–24. https://doi.org/10.26378/RNLAEL918248

García, I., & Pena, M. I. (2011). Machine translation-assisted language learning: Writing for beginners. *Computer Assisted Language Learning*, *24*(5), 471–487. https://doi.org/10.1080/09588221.2011.582687

Kliffer, M. (2005). An experiment in MT post-editing by a class of intermediate/advanced French majors. In *Proceedings of EAMT 10th Annual Conference* (pp. 160–165). European Association for Machine Translation. https://aclanthology.org/2005.eamt-1

Kliffer, M. (2008). Post-editing machine translation as an FSL exercise. *Porta Linguarum Revista Internacional de Didáctica de las Lenguas Extranjeras*, *9*, 53–68. https://doi.org/10.30827/Digibug.31745

Lacruz, I., Denkowski, M., & Lavie, A. (2014). Cognitive demand and cognitive effort in post-editing. In S. O'Brien, M. Simard, & L. Specia (Eds.), *Third workshop on post-editing technology and practice* (WPTP 2014) (pp. 73–84). AMTA. https://doi.org/10.1184/R1/6473261.v1

Lewis, D. (1997). Machine translation in a modern languages curriculum. *Computer Assisted Language Learning*, *10*(3), 255–271. https://doi.org/10.1080/0958822970100305

Loffler-Laurian, A. M. (1983). Pour une typologie des erreurs dans la traduction automatique. *Multilingual-Journal of Cross-Cultural and Interlanguage Communication*, *2*(2), 65–78. https://doi.org/10.1515/mult.1983.2.2.65

Loffler-Laurian, A. M. (1985). Traduction automatique et style (machine translation and style). *Babel International Journal of Translation*, *31*(2), 70–76. https://doi.org/10.1075/babel.31.2.03lof

Lommel, A., Uszkoreit, H., & Burchardt, A. (2014). Multidimensional quality metrics (MQM): A framework for declaring and describe translation quality metrics. *Revista Tradumàtica*, *12*, 455–463. https://doi.org/10.5565/rev/tradumatica.77

Mackey, A., & Marsden, E. (2015). *Advancing methodology and practice: The IRIS repository of instruments for research into second languages*. Routledge.

Niño, A. (2008). Evaluating the use of machine translation post-editing in the foreign language class. *Computer Assisted Language Learning*, *21*(1), 29–49. https://doi.org/10.1080/09588220701865482

Niño, A. (2009). Machine translation in foreign language learning: Language learners' and tutors' perceptions of its advantages and disadvantages. *ReCALL*, *21*(2), 241–258. https://doi.org/10.1017/S0958344009000172

Rico, C., Sánchez-Gijón, P., & Torres-Hostench, O. (2017). The challenge of machine translation post-editing: An academic perspective. In G. Corpas Pastor & I. Durán Muñoz (Eds.), *Trends in E-tools and resources for translators and interpreters* (pp. 203–218). Brill Rodopi. https://doi.org/10.1163/9789004351790_011

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 S. O'Brien, M. Simard, & L. Specia (Eds.), *Third workshop on post-editing technology and practice. 11th conference of the association for machine translation in the Americas* (pp. 5–19). https://aclanthology.org/2014.amta-wptp.1/

Schilperoord, J. (1996). *It's about time: Temporal aspects of cognitive processes in text production*. Rodopi.

Sheng, L. (2006). *Curso de traducción del español al chino* (2nd ed.). Foreign Language Teaching and Research Press.

Sycz-Opoń, J., & Gałuskina, K. (2017). Machine translation in the hands of trainee translators – An empirical study. *Studies in Logic, Grammar and Rhetoric*, *49*(1), 195–212. https://doi.org/10.1515/slgr-2017-0012

Zhang, H. (2016, Oct. 11–12). *Los errores de la traducción automática Chino > Español* [Paper presentation]. The 2nd International T3L Conference: Tradumatics, Translation Technologies & Localisation, Barcelona, Spain.

Zhang, H. (2017, March 12–16). *The use of machine translation to test Chinese students' knowledge of Spanish* [Paper presentation]. LICTRA 2017 Translation 4.0: Intracultural and Intralinguistic Features in the Subtitling Process, Leipzig, Germany.

Appendix

Sentences to be Post-edited by Participants [Source sentences from Sheng (2006)]

SENTENCE (L1)		MACHINE TRANSLATION	SOLUTIONS
1	敬请光临。此致敬礼。 [Contexto: final de una carta formal 邮件结]	Por favor. Ven. Atentamente.	*Error type*: REGISTRO FORMAL *Solution*: Confiamos en que podremos contar con su presencia. Atentamente.
2	以下简称”香蕉”。	En lo sucesivo, "banana".	*Error type*: OMISIÓN *Solution*: En lo sucesivo, llamado "banana"
3	在此意义上，两国领导人重申，对于秘鲁和中国来说，加强太平洋领域作为下个世纪最具活力的地区的地位是其必要的目标。	En este sentido, los dos estadistas reiteraron que para el Perú y China, fortalecer la consolidación del próximo siglo como la región mayor dinámica de la cuenca del Pacífico es su objetivo necesario.	*Error type*: ORDEN DE PALABRAS *Solution*: En tal sentido, los dos estadistas reiteraron que para el Perú y China, es un objetivo necesario la consolidación de la zona de la cuenca del Pacífico como la de mayor dinamismo para el próximo siglo.
4	大学校长	Presidente de la Universidad	*Error type*: NOMBRE OFICIAL *Solution*: Rector de la Universidad
5	一般来说，当年的酒或者年份少的酒价格低一些。	En general, el precio del vino o vinos de baja calidad es menor.	*Error type*: PRECISIÓN *Solution*: En general, los vinos del año o vinos jóvenes, son más baratos.

6	这就是三音石：站立于第一块石板上击掌，可听到回音一声；于第二块石板上击掌，可听到回音两声；于第三块石板上击掌，可听到回音三声。	Esta es la piedra de tres tonos: de pie en la primera placa de piedra, puede escuchar el eco, en la segunda, puede escuchar dos ecos, en la tercera, puede escuchar el eco tres veces.	*Error type*: OMISIÓN *Solution*: Esta es la piedra de tres tonos: de pie en la primera placa de piedra, (y) si da una palmada (palma), puede escuchar un eco; si se coloca en la segunda piedra y hace lo mismo, puede escuchar dos ecos; de la misma manera, en la tercera piedra, puede escuchar el eco tres veces.
7	御花园在出口附近，是故宫里最大的花园。	El Jardín Real está cerca de la salida y es el jardín más grande de la Ciudad Prohibida.	*Error type*: NOMBRE OFICIAL *Solution*: El Jardín Imperial está cerca de la salida y es el jardín más grande de la Ciudad Prohibida.
8	纪念活动闭幕式将于周五晚七点在秘鲁天主教大学文化中心举行，届时将由特邀嘉宾埃弗赖因·克里斯托博士做讲座。	La ceremonia de clausura de la conmemoración se llevará a cabo a las 7 pm el viernes en el Centro Cultural de la Universidad Católica, donde el invitado especial Dr. Evrein Cristo dará una conferencia.	*(There are no errors in this sentence)*
9	过去十年深圳经济一直维持两位数的增速，2017年地区生产总值（GRP）同比增长13.8%，在中国南部地区城市中心经济总量名列第一。	En los últimos diez años, la economía de Shenzhen ha mantenido una tasa de crecimiento de dos dígitos. En 2017, el Producto Bruto Regional (GRP) aumentó un 13,8% con respecto al mismo período del año pasado, ocupando el primer lugar en la producción económica total de los centros urbanos en el sur de China.	*Error type*: PREPOSICIÓN *Solution*: En los últimos diez años, la economía de Shenzhen ha mantenido una tasa de crecimiento de dos dígitos. En 2017, el Producto Bruto Regional (GRP) aumentó en un 13,8% con respecto al mismo período del año pasado, ocupando el primer lugar en la producción económica total de los centros urbanos en el sur de China.

10	遵照这一原则，成年人在行使其权利时不得在任何时候，任何情况下限制儿童权利的行使。	[Tested proposal]: De acuerdo con este principio, los adultos no podrán, en cualquier momento, limitar los derechos del niño en cualquier circunstancia el ejercicio de sus derechos. [New MT proposal]: De acuerdo con este principio, los adultos no deben ejercer los derechos de los niños en ningún momento y bajo ninguna circunstancia en el ejercicio de sus derechos.	*Error type*: PRECISIÓN *Solution*: [Tested proposal]: De acuerdo con este principio, los adultos no podrán, en ningún momento, limitar los derechos del niño en ninguna circunstancia el ejercicio de sus derechos. [New proposal]: De acuerdo con este principio, los adultos no deben limitar los derechos de los niños ni su ejercicio en ningún momento y bajo ninguna circunstancia.

About the Authors

Hong Zhang, PhD, is a lecturer of Teaching Spanish as a Second Language at the College of International Studies of Yangzhou University in China. Her research interests include machine translation post-editing and foreign language learning.

E-mail: zhhong2020@126.com

Olga Torres-Hostench, PhD, is a lecturer at the Faculty of Translation and Interpreting at the Autonomous University of Barcelona. Her research interests are machine translation post-editing and translation skills acquisition. At present, she teaches specialized translation and multimedia localization.

E-mail: olga.torres.hostench@uab.cat