Correlations between productivity and quality when post-editing in a professional context

Ana Guerberof Arenas
ana.guerberof@gmail

Abstract. This article presents results on the correlation between machine-translated and fuzzy matches segments in terms of productivity and final quality in the context of a localization project. In order to explore these two aspects, we set up an experiment with a group of twenty four professional translators using an online post-editing tool and a customized Moses machine translation engine with a BLEU score of 0.60. The translators were asked to translate from English to Spanish, working on no-match, machine-translated and translation memory segments from the 85-94 percent value, using a post-editing tool, without actually knowing if the segment came from machine translation or from translation memory. The texts were corrected by three professional reviewers to assess the final quality of the assignment. The findings suggest that translators have higher productivity and quality when using machine-translated output than when translating without it, and that this productivity and quality is not significantly different from the values obtained when processing fuzzy matches from translation memories in the range 85-94 percent.

Key words: post-editing; productivity; quality; machine translation; translation memory.

1 Introduction

Translation Memories (TM) have been used extensively in the localization industry for the past twenty years, while Machine Translation (MT) has been incorporated in the mainstream localization workflow for a shorter period of time even though MT pre-dated TM. Increasingly, however, MT segments are included together with fuzzy matches within the standard localization workflow. This development has triggered questions on the possible benefits that would help to reduce the translation cost while maintaining the same level of quality. We are presenting in this article the results from a project carried out as part of a doctoral thesis (Guerberof 2012) that was seeking to explore precisely these aspects: productivity and quality when working with MT output. We studied the work of professional translators by establishing correlations between MT and TM segments that could help us better understand productivity metrics and establish a more solid ground for pricing MT segments. At the same time, it was necessary to test, if by using different inputs, MT or TM, the quality was in any way altered. Moreover, we needed to compare these two "methods" to the translators' speed and quality delivered when having no proposal (using these figures as a baseline for comparison). We will focus here on the analysis of these two aspects, although related aspects such as translators' experience, reviewers' processes and edit distance of the reviewed segments were also explored as part of this doctoral thesis.

2 Key concepts

A fuzzy match is a partial proposal that the TM system proposes if an exact match is not available, an exact correspondence between the old and new source, also called 100 percent match. Different levels of fuzzy matches are calculated with an algorithm that uses character string similarities and establishes partial correspondences between the source sentences based on syntactic structures. However, different tools use different algorithms. The fuzzy match value is normally expressed as a percentage. This percentage represents the characters that were translated with a less than perfect translation memory match (100 percent). Therefore, an 85 percent match, for example, is considered a high percentage match, that is, the translation proposed is deemed to be very close to the new source text. In this paper, the fuzzy matches refer to the ones generated by SDL Trados 2007. For consistency purposes we have used a similar terminology for the MT output generated by a trained Moses engine (see section on Methodology) and we have named these segments *MT match*.

3 Related Work

The integration of machine translation and the task of post-editing in the localization industry workflow have motivated an array of studies in academia and commercial settings. Although this number has increased in recent years and there are many that are relevant to other areas of machine translation or translation studies research, we can only make reference here to those that have most significantly influenced this particular article.

In an academic setting, Sharon O'Brien has studied various aspects of post-editing: post-editors' profiles and course contents for post-editing (2002), and the correlation between post-editing effort and machine translatability, suggesting a new methodology for measuring source-text difficulty and cognitive effort (2006a). Very relevant to this study, as it served as the inspiration to the thesis this article stems from, is her research on eye tracking and translation memory matches (2006b). In this paper, O'Brien shows that as the fuzzy match value decreases, the cognitive load increases and that MT matches appear to be equivalent to 80-90 percent fuzzy matches (SDL Translator's Workbench) in TMs in terms of cognitive load. Although, there are many differences between O'Brien's study and ours in terms of goals, number of subjects, language combinations, quality evaluation, among others, we found that the idea of comparing the values between fuzzy matches and MT matches to explore in depth translators' usage of MT output was innovative and illustrative (for practitioners) and we have used it as a key concept in our research. With respect to quality in post-edited material, there are an increasing number of studies that show, contrary to a more "popular belief" amongst translation practitioners, that quality when applying MT output and then post-editing is not lower than when using only human translation and that, on occasions, MT can serve as an aid. This is the case in Offersgaard et al. (2008), Fiederer and O'Brien (2009), O'Brien (2011), Carl et al. (2011), García (2010) and (2011), and De Sutter and Depraetere (2012). Each of these experiments might include different language combinations, use a different MT engine, or include a different number of participants, but overall they help to clarify what is to be expected when MT is introduced in a translation environment. Our study is innovative in comparison to these because it uses a higher number of professional translators but it shares the goal of creating more data on the productivity and quality that can be expected when using MT in a localization environment. De Almeida and O'Brien (2010) explore the possible correlation between post-editing performance and years of translation experience to find that the most experienced are the fastest post-editors and that they make the higher number of essential changes. However, the results also show that the translators with more experience make more preferential changes. The methodology for this study is interesting for us, especially the edit analysis divided into different categories using the LISA QA model. Although we used a higher number of participants in a different language combination, we share the use of a similar quality framework to analyze the number and type of errors. Tatsumi (2010) studies the speed of nine professional English to Japanese post-editors and analyzes the amount of editing made during the process, as well as the influence of the source text on speed and type of edits. The results show that the amount of editing moderately correlates with post-editing speed. This means that post-editors with different speeds might make the same number of edits. Also, and in relation to our work, the editing speed of fuzzy matches (75-99 percent matches) is not significantly different from MT matches in terms of speed. We feel that the sample of participants is small if statistical conclusions in terms of speed are to be drawn. However, this is a very thorough and solid study that addresses important questions in post-editing through the behavior of post-editors in a commercial environment, and it sheds some light on the nature of changes and their correlation to speed.

In a commercial setting, Flournoy and Duran (2009) find that post-editors are faster than translators unassisted by technology although they observed differences depending on the set of files used in a live project. They also found that novice translators were more prone to accept MT proposals too readily leading to lower final quality. It is difficult to know how these figures are calculated in the context of a live project. There are, however, interesting comments in this paper in relation to the following topics: MT quality and post-editing speed vary significantly between files; MT quality relation to the quality of the source text; integration of MT within a Globalization Management System (GMS); implementation of feedback from translators; and the profile needed for post-editors. Plitt and Masselot from Autodesk (2010) report on a very interesting productivity test of statistical machine translation post-editing. It is significant for us for two reasons: it is similar to our own research (productivity and quality) both in methodology and findings, but furthermore the researchers work in a commercial setting. This is important because it signals, in our view, a slight change in how post-editing and post-editing “experiments” are being viewed and carried out in the localization industry. Since the origin is commercial, the study is more focused on the productivity of the post-editing task, and this is understandable, since an increase in productivity can mean a reduction in costs. The advantage of these types of studies is that they are carried out in an environment “similar” (with similar texts) to that of professional translators. They set up an experiment with 12 participants into four different languages (French, German, Italian and Spanish) and they found a high variance across translators; MT allowed translators to work faster but the percentages varied from 20 to 131 percent. They also find that the benefits of MT are greater for slower translators than for faster ones. Interestingly, the number of changes made is not necessarily lower for faster translators; in fact, the fastest translator has the highest number of changes. Therefore, no correlation is found between edit distance and throughput. The Autodesk QA team reports, after reviewing all proposals, a higher number of errors for translation jobs than for post-edited jobs in all the languages tested. As a follow up to this study, Autodesk (2011) has published on their website the results of a two-day translation and post-editing test carried out with 37 participants from English into Chinese, Japanese, Polish, Portuguese, German, Italian, Korean, Spanish and French. The results show that for all languages and participants, post-editing productivity is higher than translation productivity. The productivity increase is different for different languages, for example Chinese is the lowest (with 42 percent) and French is the highest (with 131 percent). Spanish has a 117-percent increase. Experience, especially in post-editing, is considered to be the most important factor in productivity. When comparing with fuzzy matches (of all categories including below 50 percent matches) with MT in those languages with best MT output (Chinese, Polish, German, Spanish, French and Italian), MT is more productive globally, and as productive as the matches in the 85-94 range in particular. A blind final translation quality evaluation was carried out for both types of translations and the results reveal that reviewers cannot tell the difference between post-edited text and full human translation, and there is not a pattern that suggests loss of quality.

It is important to mention that a pilot project, the details of which are described in Guerberof 2008, was carried out with eight subjects. The results showed that post-editing MT segments was in fact faster on average (mean value considered) than editing TM segments. Nonetheless, the data dispersion was very high among participants, suggesting high subject dependency. When quality was analyzed, the number of errors in TM segments was higher than in MT segments (by 91 percent) and than in segments translated without any aid (by 141 percent).

The number of errors in the MT and human translated segments was quite close, despite being slightly higher in MT. The sample of eight participants was a highly limiting factor. It was necessary to explore further the relationship between productivity and quality with a greater number of participants.

4 Methodology

As a result of the pilot study (Guerberof 2008) we were looking to test if the time invested in post-editing machine-translated text would correspond to the time invested in editing fuzzy-matched text corresponding to the 85-94 percent range. Moreover, we wanted to know if any increase in speed using MT technology would have any impact on the resulting quality of the segments.

With this objective in mind, a trained Moses (Koehn et al. 2007) engine was used to create the MT output. In order to train the engine, we used a translation memory (TM) and three glossaries. The TM used came from a supply chain management provider (IT domain) and it had 173,255 segments and approximately 1,970,800 words (English source). The Excel glossaries contained 610 entries and 94 entries of core terminology; the XML file contained 9,106 entries (software strings in xml format). The resulting output obtained a BLEU score (Papineni et al. 2002) of 0.60 and a human evaluation score of 4.5 out of 5 points (5 being Excellent and 1 being Very Poor according to predefined criteria). The file set used in the actual project was a new set of strings for the help system and user interface from the same customer and therefore different from the parallel data used to train the engine. To prepare the file, we pre-translated the new files with the TM in order to obtain fuzzy matches using the option Pre-translate in Trados. We selected the segment pairs together with the corresponding fuzzy match level. We used the fuzzy match level 85-94% for the Fuzzy match option. We took the source segments that fell in the 0-50 percent range and we machine translated them. We deleted all duplicates and all segments with length below 4 and above 26 words. The three-word segments, without context, can generate a lot of doubts during translation, whereas in the case of the segments over 26 words, there were very few in the corpus and we did not have sufficient material to replicate them in all categories (No, Fuzzy and MT match). We replaced all tags with place holders ({ph}) and then we applied a corpus sampler. The sampler creates a histogram with segment lengths of the original corpus and it applies the same length distribution to the sample. We applied this same pattern to each category (No match, Fuzzy match and MT match) so that the distribution in each category was balanced. The final file set contained 2,124 words in 149 segments distributed as follows: No match, 749 words, MT match (the output), 757 words and Fuzzy match, 618 words from the 85 to 94 percent range. The 24 translators received the same set of segments and they had the task to translate the No match and edit the MT and Fuzzy matches (they were not aware of the origin of each proposal). Three initial segments (one of each category) were included for translators to practice on and become familiar with the tool, the glossary and the instructions. These were not included when measuring productivity or quality. Figure 1 shows the workflow for creating the dataset.

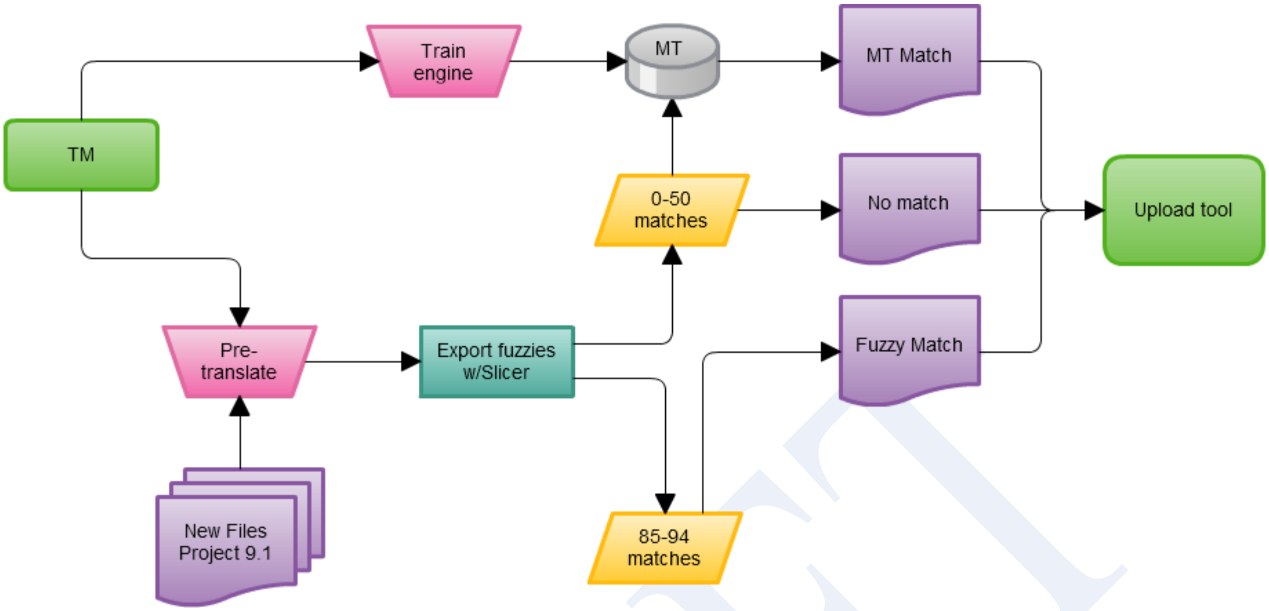
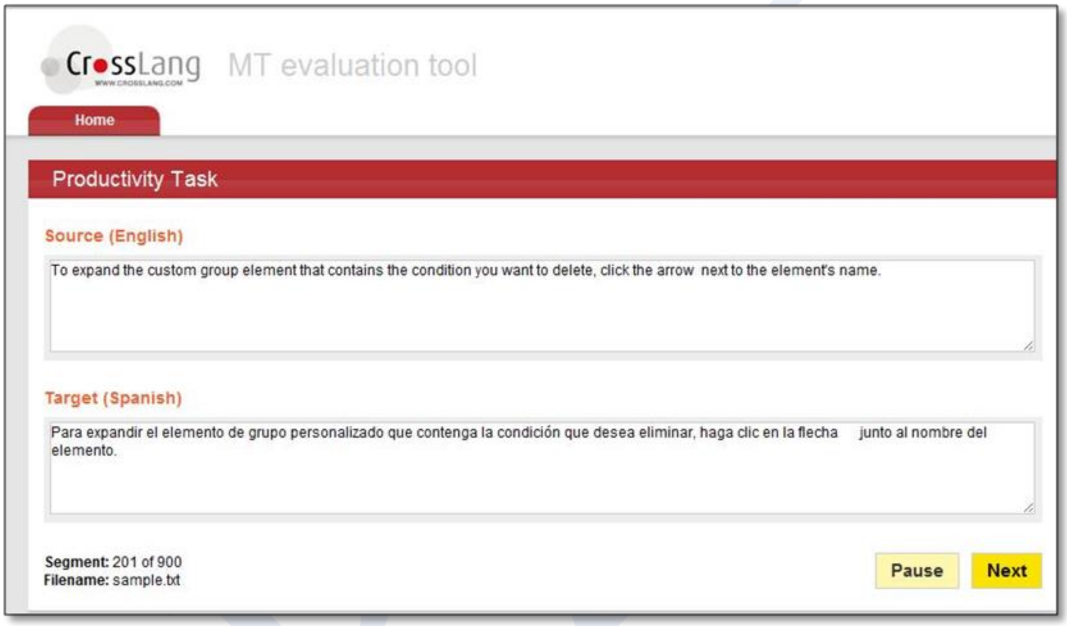


Figure 1: Dataset creation

For this experiment, we selected 24 professional translators from English to Spanish from those approved in a large Language Service Provider (Pactera International) database. The translators are approved through a selection process that involves CV selection, testing and sample testing within live projects. Since post-editing is a relatively new task for language service providers and the number of post-editors available is limited, the Vendor Management team usually contacts translators with or without experience when a new post-editing project arrives. Therefore, we followed exactly the same criteria: no post-editing experience was required a priori but translators with post-editing experience were not to be discarded. The three professional reviewers had to have at least three years' experience in localization (software, help and/or documentation) and in Computer Aided Translation Tools (SDL Trados, Déjà Vu, MemoQ or similar tools). Familiarity with tools is an indication of familiarity with translation memories and post-editing, and reviewers needed to be aware of the environment translators were working in to assess the texts produced by them. The reviewers should also have at least six months' working experience in MT post-editing and in Business Intelligence software translation. These criteria were applied so reviewers were sufficiently familiar with the task at hand so as to evaluate the work done by the translators without introducing unnecessary changes and, at the same time, have sufficient technical knowledge to perform the task at the standard review speed.

The translators and reviewers available were informed about the nature of the project, the rate (the fee agreed was the standard full rate per word for this language combination) and the time frame, as in a standard localization project. They were asked to sign a Research Participant Release Form approved by the University Rovira i Virgili. The form clearly stated that their participation was voluntary and that they granted permission for the evaluation of the data without identifiable information in regard to their name.

In order to measure the actual times and output produced by the post-editor we used a post-editing tool created by CrossLang as shown in Figure 2. The post-editors could connect online and translate or post-edit the proposed segments of text without knowing their origin (MT, TM or No match segments) and the tool measured the time taken in seconds for each segment. The Source window contained the source text in English, and the Target window contained either a blank text box or a proposed text in Spanish. The Spanish text was either an MT or TM segment. The segments were presented to the translators randomly due to the way in which the data set was created. However, this is not unusual in a "real" work environment where translators are asked to process only certain fuzzy match segments in a non-linear way. Once the translators were finished, we exported all the data from translators: translator name, segment ID, source text, MT target text, post-edited text, TM data, MT/PE difference (%), segment length, fuzzy match level, post-edit time in ms, and segment origin.



Productivity Task

Source (English)

To expand the custom group element that contains the condition you want to delete, click the arrow next to the element's name.

Target (Spanish)

Para expandir el elemento de grupo personalizado que contenga la condición que desea eliminar, haga clic en la flecha junto al nombre del elemento.

Segment: 201 of 900
Filename: sample.txt

Figure 2: Screen-shot of post-editing user interface used

The translators received a set of instructions by e-mail explaining exactly the steps they needed to take to translate, edit and post-edit the segments. The instructions included how to interact with the tool, carry out the assignment, translate software options, follow specific guidelines on style, and how to use the Excel glossary provided (this glossary contained core terms provided by the customer). They were also instructed about the quality: publishable quality, meaning full accuracy and no mistranslations with regards to the English text, compliance to Spanish language rules of grammar and spelling, compliance to the terminology following the glossary provided, and compliance to style according to the style guidelines provided. Because the tool did not allow the translators to review the segments once they had clicked the Next button, they were reminded in the guidelines that they had to review each segment fully before going to the next segment.

The final output was then evaluated by the three professional reviewers, who registered the errors using the LISA QA model. This is a model created by the Localization Industry Standards Association (LISA) that it is widely used in the localization industry to evaluate the quality of translated material by classifying errors found in the target text. The reviewers had to read the source text, the proposed translation and then they had to read the post-edited target text to make sure it reflected the changes that would match the source text. They were also advised not to insert any preferential change (to only correct errors), to count the same error across a translation only once, to make sure the translators followed the glossary provided, and to be consistent across the 24 translations. Finally, they were given some clarifications on error typology according to the LISA QA model. They were reminded that their ability to spot errors was not being judged, so that they should only mark the errors found in the final translation that they were absolutely sure of.

5 Results on productivity

The processing times (in milliseconds) for all 24 translators were recorded by the tool during the assignment and we are presenting here the processing speed (words per minute) per translator. Figure 3 shows the processing speed according to the different match categories for all segments processed in the assignment:

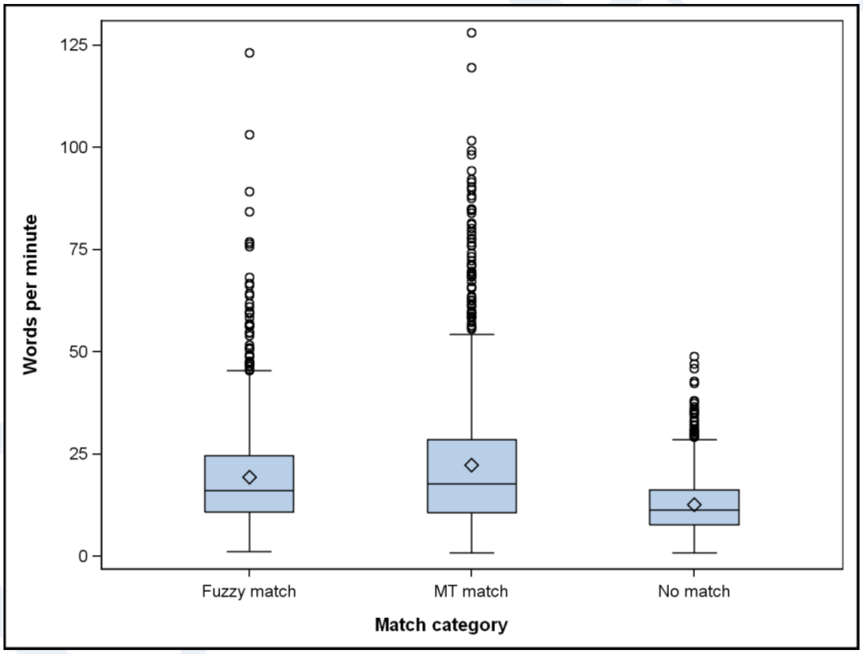


Figure 3: Global processing speed according to match category

The box-and-whisker diagram shows the distribution of the dataset from 0 to 125 words per minute according to quartiles for the three types of matches. The actual blue/grey box includes data from the first to the third quartile (50 percent of the data). The median value is represented by a black line. The upper whisker represents the data from the third quartile to the maximum value and the lower whisker represents the data from the first quartile to minimum value. The mean value is represented by a diamond and the outliers are presented by circles. Outliers are extreme values that deviate from the rest of the samples.

We can see in Figure 3 that the most words processed per minute on average are in the MT match category, followed by the Fuzzy match, and lastly by the No match category. There are, however, more homogeneous figures for the No match than for the other two categories, where the data have greater dispersion, and there are more outliers. Table 1 shows the descriptive analysis of the data.

Match category	**Number of segments**	**Mean**	**Median**	**SD**	**Min**	**Max**
Fuzzy match	1197	19.33	16.13	12.67	1.15	123.12
MT match	1176	22.34	17.71	17.27	0.89	128.11
No match	1198	12.64	11.35	6.97	0.82	48.91

Table 1: Descriptive statistics for global processing speed (words per minute)

The mean and median values, as in Figure 3, are higher for MT match, followed very closely by Fuzzy match and then No match, but there is less deviation when processing the latter, showing a more homogeneous processing speed when translators work without a translation proposal. This could be due to the nature of particular segments and to the difference in the proposals from MT and Fuzzy matches. Some segments might require more editing (hence the minimum values 0.15 or 0.89 words per minute), while others might require only reading if the quality of the output is high. For example, an MT match might require no post-editing (we see a maximum value of 128.11 words per minute), a Fuzzy match might require a slight edit (we see a maximum value of 123.3 words per minute), but a No match will always require translators to translate the segment fully, hence processing fewer words per minute in comparison to the other two categories (the maximum value in this category is 48.91 words per minute).

We can see that these results show that MT increases translators' speed, in line with previous research (Guerberof 2008, Offersgaard et al. 2008, Masselot and Plitt 2010, De Almeida and O'Brien 2010). Tatsumi (2010) also showed very similar results when studying the English to Japanese language combination: she reported 22.38 words per minute as a mean, and our mean value is very close, at 22.34 words per minute.

In order to see if the correlation between MT and Fuzzy matches is significant we applied a linear regression model with repeated measures, taking as the response variable *logarithm of Words per minute*. For this variable, statistically significant differences are observed (F=239.96 and p <0.0001) between the three categories of matches: Fuzzy, MT and No match. From this model, estimates of the mean values of the variable *logarithm of Words per minute* according to the *Match category* have been obtained. To better interpret the results, Table 2 shows the corresponding estimates expressed with the same units as the original variable *Words per minute* according to the match category and the corresponding confidence intervals of 95 percent (Lower and Upper).

Match category	**Estimated mean**	**Lower**	**Upper**
Fuzzy match	15.95	13.79	18.45
MT match	17.21	14.88	19.91
No match	10.79	9.33	12.49

Table 2: Estimated global words per minute

We can observe in Table 2 that the lower and upper estimates overlap in the Fuzzy (13.79 to 18.45 words per minute) and MT match categories (14.88 to 19.91 words per minute), while the same estimates for No match do not (9.33 to 12.49 words per minute). Therefore, we can infer that Fuzzy match and MT match values do not show statistically significant differences while the No match does. In other words, the translators were faster when processing the Fuzzy and MT matches than when translating on their own (No match) and, furthermore, the processing speeds for MT and Fuzzy match segments in the 85-94 percent range are not significantly different.

This is in line with our previous work (Guerberof 2008), where we observed that processing MT matches was faster than processing Fuzzy matches from the 80-90 percent range if the mean value was considered. This followed the study mentioned earlier by O'Brien (2006b) that had indicated a correlation between these two categories. Tatsumi (2010) shows a comparison of TM and MT matches in her study (although she uses a low volume of fuzzy matches in comparison to the volume of MT output used) and she concludes that the speed for processing MT matches lies within the fuzzy match range of editing 75 percent TM matches (English to Japanese). However, it needs to be taken into account that when looking at other studies, the language combinations and engines are different and thus it is difficult to directly compare results.

In the commercial sector, JABA Translations, a Portuguese and Spanish language service provider, has reported a correlation similar to the one obtained here (Asia Online 2012) in the English to Portuguese language combination and using a customized Asia Online engine, although we do not currently have any published data on how the experiment was performed. Autodesk (2011) has published a study online carried out in several languages where "MT was roughly as productive as working from matches in the 85-94% category. Note however that this varies significantly across languages". Our results are not that disparate. This could be an indication of the high correlation level possible between MT matches and high Fuzzy matches if the MT engine is trained with sufficient data, both in terms of high volume (quantity) and high quality data.

5.1Productivity gain

We want to explore the percentage of gain that translators have when processing MT or Fuzzy matches with respect to their No match processing speed. In order to do this, we model the data using a linear regression model with repeated measures. We present it here using the original variable *Words per minute* according to MT match for better understanding. We have added a column to the right expressing the percentage of gain. This gain has been calculated taking only the estimated mean value (processing speed) of No match and comparing it with the estimated mean value of the other two categories (Fuzzy and MT match). Therefore it does not mean that it is applicable to every single segment processed by translators. We have added as well a Time savings column expressing the percentage of time saved for each translator considering the time invested in translating an amount of words designated as 1 at a No match speed, minus the time invested in doing the same amount of words at a MT or TM speed. We have used the following formula (as in Plitt and Masselot 2010), $x = 1 - \frac{1}{1+y}$ where x is the time saved and y is the productivity gain (Fuzzy and MT match). For example for Translator 1, the Fuzzy match time saving is: $0.36 = 1 - \frac{1}{1+0.55}$. Table 3 shows all these results.

Translator	Match	Estimated mean	Lower CI 95%	Upper CI 95%	Estimated productivity gain	Time savings
TR1	Fuzzy	21.49	18.36	25.17	**55.16%**	36%
TR1	MT	21.03	17.93	24.67	51.81%	34%
TR1	NM	13.85	11.83	16.22	.	
TR2	Fuzzy	12.57	11.17	14.14	38.76%	28%
TR2	MT	13.71	12.18	15.44	51.37%	34%
TR2	NM	9.06	8.05	10.19	.	
TR3	Fuzzy	9.11	7.64	10.85	86.49%	46%
TR3	MT	9.61	8.05	11.47	96.89%	49%
TR3	NM	4.88	4.10	5.82	.	
TR4	Fuzzy	9.44	8.08	11.03	32.89%	25%
TR4	MT	9.77	8.36	11.42	37.60%	27%
TR4	NM	7.10	6.08	8.30	.	
TR5	Fuzzy	24.09	20.92	27.74	47.78%	32%
TR5	MT	27.69	24.01	31.93	69.86%	41%
TR5	NM	16.30	14.16	18.77	.	
TR6	Fuzzy	14.21	11.96	16.88	96.48%	49%
TR6	MT	18.07	15.18	21.51	149.9%	60%
TR6	NM	7.23	6.09	8.59	.	
TR7	Fuzzy	18.49	15.47	22.11	38.93%	28%
TR7	MT	20.31	16.96	24.33	52.61%	34%
TR7	NM	13.31	11.13	15.91	.	
TR8	Fuzzy	19.02	16.45	22.00	47.82%	32%
TR8	MT	22.98	19.85	26.62	78.58%	44%
TR8	NM	12.87	11.13	14.88	.	
TR9	Fuzzy	12.23	10.21	14.65	**41.24%**	29%
TR9	MT	10.81	9.04	12.92	24.75%	20%
TR9	NM	8.66	7.25	10.35	.	
TR10	Fuzzy	20.72	17.59	24.40	47.63%	32%
TR10	MT	23.85	20.22	28.13	69.91%	41%
TR10	NM	14.03	11.92	16.53	.	
TR11	Fuzzy	7.98	6.82	9.33	18.21%	15%
TR11	MT	10.19	8.70	11.93	51.01%	34%
TR11	NM	6.75	5.77	7.89	.	
TR12	Fuzzy	12.12	10.16	14.46	**80.13%**	44%
TR12	MT	11.19	9.37	13.37	66.34%	40%
TR12	NM	6.73	5.64	8.03	.	
TR13	Fuzzy	28.34	23.98	33.48	40.78%	29%
TR13	MT	31.66	26.75	37.47	57.29%	36%
TR13	NM	20.13	17.03	23.78	.	
TR14	Fuzzy	10.08	8.50	11.96	45.49%	31%
TR14	MT	10.61	8.92	12.61	53.08%	35%
TR14	NM	6.93	5.84	8.22	.	
TR15	Fuzzy	26.25	23.01	29.95	57.78%	37%
TR15	MT	26.31	23.03	30.06	58.11%	37%
TR15	NM	16.64	14.58	18.98	.	
TR16	Fuzzy	16.08	14.14	18.28	**56.86%**	36%
TR16	MT	15.97	14.03	18.18	55.86%	36%
TR16	NM	10.25	9.01	11.65	.	
TR17	Fuzzy	14.55	12.52	16.90	**50.48%**	34%
TR17	MT	14.03	12.06	16.33	45.14%	31%
TR17	NM	9.67	8.32	11.23	.	
TR18	Fuzzy	21.11	18.11	24.61	65.02%	39%
TR18	MT	24.31	20.82	28.38	90.05%	47%
TR18	NM	12.79	10.97	14.91	.	

Translator	Match	Estimated mean	Lower CI 95%	Upper CI 95%	Estimated productivity gain	Time savings
TR19	Fuzzy	21.10	17.94	24.83	34.23%	26%
TR19	MT	26.73	22.68	31.51	70.03%	41%
TR19	NM	15.72	13.36	18.50	.	
TR20	Fuzzy	19.35	17.22	21.75	23.46%	19%
TR20	MT	20.19	17.95	22.72	28.81%	22%
TR20	NM	15.67	13.95	17.62	.	
TR21	Fuzzy	17.55	14.86	20.74	**63.04%**	39%
TR21	MT	16.94	14.31	20.05	57.36%	36%
TR21	NM	10.77	9.11	12.72	.	
TR22	Fuzzy	12.79	10.62	15.41	46.47%	32%
TR22	MT	14.70	12.18	17.74	68.34%	41%
TR22	NM	8.73	7.25	10.52	.	
TR23	Fuzzy	20.16	18.20	22.33	6.21%	6%
TR23	MT	22.89	20.65	25.38	20.62%	17%
TR23	NM	18.98	17.14	21.02	.	
TR24	Fuzzy	15.71	13.56	18.19	56.74%	36%
TR24	MT	16.60	14.31	19.25	65.60%	40%
TR24	NM	10.02	8.65	11.61	.	

Table 3 Estimated WPM and productivity gain per individual translator

All 24 translators have some productivity gain when using MT and Fuzzy matches when estimated mean values are compared. Six translators out of these - 1, 9, 12, 16, 17 and 21 (highlighted in bold in the table) - showed a higher gain when using Fuzzy matches, and consequently 18 translators showed a higher gain when using MT matches.

It is also interesting to note that the productivity gain for MT matches ranges from 20.62 (TR23) to 149.90 percent (TR6), and in the case of Fuzzy matches from 6.21 (TR23) to 96.48 percent (TR6). As we can see, in spite of an increase in both categories, there are different degrees for individual translators. One translator might increase productivity by 150 percent while another might do it by only 20 percent. Further, it can be seen that some translators (for example, translator 9 or 23) might have a higher estimated mean value in MT and Fuzzy, hence the productivity gain, but the confidence intervals show similar values in the three categories, and this might indicate that this gain is not so clear with some translators.

6 Results on quality

Quality is measured according to the number of errors present in the final target text. The reviewers classified the errors according to the different segment categories (No match, Fuzzy match and MT match), using the LISA QA standard. The focus is on the number and classification of errors, as the aim of this study is not to establish if a particular translator offers good or poor performance but if the number and type of errors is conditioned by the use of a translation aid and therefore if the errors have an impact on the overall productivity of the translation. That is, we seek to investigate whether the time saved using MT (or TM) does not mean additional time in order to fix errors at a later stage in the localization process. The reviewers were informed about the type of match because they were requested to mark any overcorrection (those edits performed by translators that did not necessarily address an error in the MT output or TM Fuzzy match). However, the reviewers marked very few overcorrections to produce meaningful data. More information about overcorrections can be found in the complete thesis (Guerberof 2012).

6.1 Errors at segment level

In order to see the number of segments that contain errors and those that did not, the data from the segment database were used and the following values obtained as shown in Table 4.

Match	**No errors**		**Errors**	
	Number of segments	**%**	**N**	**%**
Fuzzy match	2946	81.83	654	18.17
MT match	2889	81.89	639	18.11
No match	2618	72.72	982	27.28
All	8453	78.79	2275	21.21

Table 4: Error indicator per category

Almost 80 percent of all segments reviewed did not contain any errors (78.79 percent) and 21 needed some kind of correction. The Fuzzy and MT matches have almost identical percentages of segments that were corrected by reviewers, 18.17 and 18.11 percent respectively. Even the No match category, which contains more segments edited (27.28 percent), still has 72.72 percent of segments that did not require any change. Of course, the quality of a translation is not measured only in terms of those segments that are not corrected but according to the type and number of errors as well. Nevertheless, judging from these results, translators delivered quite a high percentage of error-free segments according to the reviewers. Attention should be drawn to the fact that although this task was reviewed by three professional translators with experience, this does not mean that the reviewers cannot make mistakes or insert preferential changes. We measured the agreement between reviewers by applying a Kappa coefficient per Match category and Reviewer. We observed that for the No match category there was agreement between the three reviewers, 0.61 to 0.75 (p=<.0001). However, for Fuzzy and MT match, the Kappa coefficient is between values that show either no agreement (-0.02 ≈ 0) or a weak agreement (0.43). More information on reviewers' agreement can be found in the complete thesis (Guerberof 2012).

We modeled the data with repeated measures, taking *Error indicator* as the response variable, to observe the possibility of making an error in a segment depending on the category. Statistically significant differences are observed in the proportion of errors in the three different types of translations Fuzzy match, MT match and No match (F=62.15, p<0.0001). From this model, the estimated odds of "making a mistake" while translating, editing or post-editing are calculated. Table 5 shows this value with the corresponding confidence intervals of 95 percent.

Match	**Odds error**	**Lower**	**Upper**
Fuzzy match	0.20	0.24	0.17
MT match	0.20	0.24	0.17
No match	0.35	0.41	0.30

Table 5: Odds of "making a mistake"

In Fuzzy match, the estimated odds of "making a mistake" are 0.20, that is, the probability of making a mistake is 0.2 times the probability of not making a mistake. In MT match, the estimated odds are also 0.2 times. In No match, the estimated odds are 0.35 times the probability of not making a mistake. This match category is the one with the highest probability of making an error. The Odds *Ratio* tells us the relation between No match and Fuzzy match and No match and MT match and its confidence levels, as shown in Table 6.

Match	Match	Odds Ratio	Lower	Upper
No match	Fuzzy match	1.75	1.56	1.96
No match	MT match	1.75	1.56	1.97

Table 6: Estimated Odds ratio per category

The Odds Ratio estimation of No match with regards Fuzzy match is 1.75, that is, the odds of "making a mistake" in No match is 1.75 times the odds of "making a mistake" in Fuzzy match. The Odds Ratio of No match with regards to the MT match is 1.75. In other words, the odds of "making a mistake" in No match is 1.75 times the odds of making a mistake in the MT match category.

6.2 Error count

There are more No match segments containing errors, and the similarities between Fuzzy and MT matches are high. It is interesting to examine the number of errors per translator and category to see if there are more errors in the No match categories. Table 7 shows the final number of errors per translator, according to match category, and the total number of errors. The errors per translator are the sum of the errors from the three reviewers. The table is sorted according to ascending total errors. Totals are highlighted in bold.

Translator	Fuzzy match	MT match	No match	Totals
Translator 15	11	9	24	**44**
Translator 20	10	16	19	**45**
Translator 08	15	20	13	**48**
Translator 12	17	18	15	**50**
Translator 06	13	20	23	**56**
Translator 23	14	14	29	**57**
Translator 05	10	19	29	**58**
Translator 11	19	23	16	**58**
Translator 17	13	15	30	**58**
Translator 16	26	16	21	**63**
Translator 02	14	15	36	**65**
Translator 09	26	11	36	**73**
Translator 24	23	22	31	**76**
Translator 21	18	23	39	**80**
Translator 14	16	20	51	**87**
Translator 01	29	23	37	**89**
Translator 04	21	22	53	**96**
Translator 22	16	42	40	**98**
Translator 07	37	24	44	**105**
Translator 03	27	41	72	**140**
Translator 19	45	37	64	**146**
Translator 18	38	43	69	**150**

Translator 13	32	53	72	**157**
Translator 10	52	30	85	**167**
Totals	542	576	948	**2066**

Table 7: Number of errors per match category and translator

Table 7 shows that all segment categories contain errors and that all translators have errors in all categories. There are a total of 2,066 errors in the final texts (this is the aggregated total from the three reviewers). A total of 948 errors are found in the No match segments, 576 in the MT match and 542 in the Fuzzy match. We nevertheless see that in 19 cases there are more errors in the No match segments than in the other two categories. In four cases, there are more errors in MT (Translators 8, 12, 11 and 22); in one case (Translator 16) there are more errors in Fuzzy match. For those translators with over 100 errors, the No match category consistently has the highest amount of errors, and there is more variance in those translators with fewer errors. We can also see that in 15 cases there are more errors in MT than in the Fuzzy match category. These results, however, are the total error numbers; they do not take into account the number of words processed in each category, which was somewhat lower for Fuzzy matches. Although it is unlikely that the result would change for the No match category taking into account the volume, it could change for the other two categories. For example, Translators 17, 2, 4 and 18, with lower number of errors in Fuzzy match in absolute values, would present fewer errors in MT match if the volume of words were factored in, that is, they would present a lower error rate per word in MT match.

Translator 15 has the lowest number of errors, followed closely by Translator 20. Translator 10 has the highest number of errors, followed by Translator 13. Translator 13 was in fact the fastest, which might indicate that this translator processed the segments very quickly but she did not fully post-edit or edit the segment. However, Translator 3, the slowest translator, has 140 errors and is placed not too far from Translator 13 in terms of errors. This might suggest that spending more time on the texts does not necessarily mean a lower number of errors. As we observed repeatedly in this study, there are striking differences between translators when processing the segments. Let us now look more in depth at the results of errors per category.

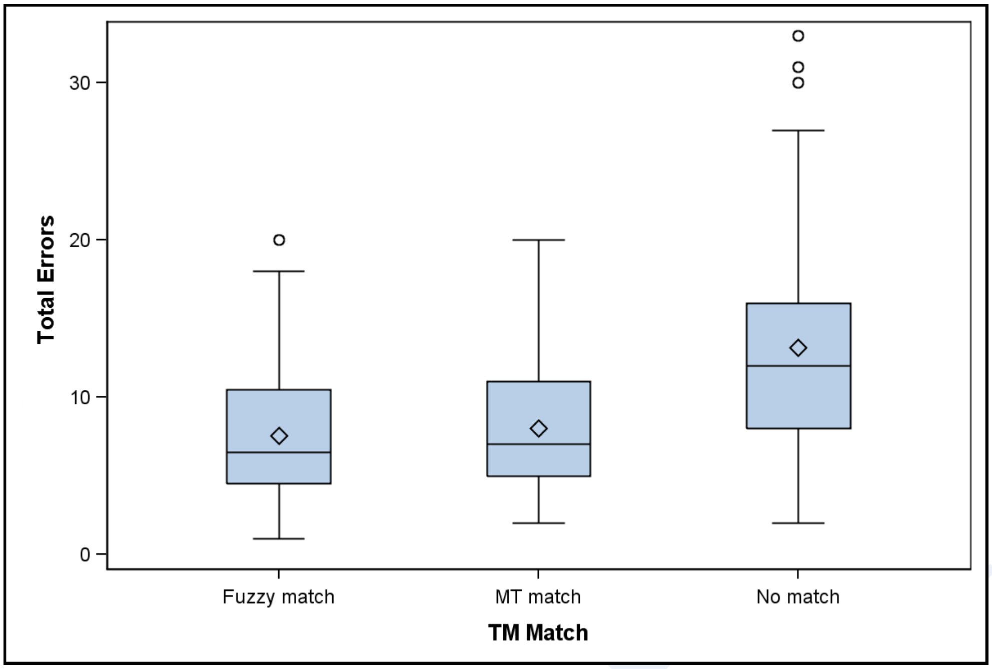


Figure 4: Total errors per translator according to match category

Figure 4 shows more errors in No match, and similar number of errors in the Fuzzy and MT match categories, although slightly lower for Fuzzy, which suggests that in this case the proposed translations helped participants to produce better quality and at a faster speed. The No match category shows more dispersion, that is, there are more differences between translators. The outliers indicate that one translator had more than 30 errors while others had fewer than 10. The ranges are smaller in the Fuzzy and MT match categories, from above 10 to around 20 errors. Table 8 shows the statistical descriptive data of the sample.

Match	N	Min	Q 1	Mean	Median	Q 3	Max	SD	Range	Q Range
Fuzzy match	72	1.00	4.50	7.53	6.50	10.50	20.00	4.11	19.00	6.00
MT match	72	2.00	5.00	8.00	7.00	11.00	20.00	4.15	18.00	6.00
No match	72	2.00	8.00	13.17	12.00	16.00	33.00	7.22	31.00	8.00

Table 8: Descriptive analysis of total errors per translator

Here again the figures for Fuzzy and MT match are very close, and those for No match stand out with higher errors, higher deviations and wider ranges. Translator 10 shows the maximum value, with 33 errors in No match and Translator 15 shows the minimum value with 1 error in Fuzzy match. The mean and median values are very close between MT match and Fuzzy match and higher for No match. In general, errors in Fuzzy and MT match are more homogenous than in No match, and this might suggest that the translators that make most errors will make more when translating from scratch (No match) than when they are using the proposed translation (Fuzzy or MT match), while translators with fewer errors have less of a difference between categories.

When analyzing the classification of these errors, we found that Language and Style were problematic areas in the No match category. Accuracy seemed to present problems in Fuzzy matches, and Mistranslations in the case of MT matches. This seems quite logical: in the case of No matches, these strings have never been reviewed (this was, in fact, the first time they were translated) while in the case of translation memories and machine translation, the segments had been "extracted" from the original translation memory. In the case of Fuzzy matches, the main changes to be made in the 85-94 range are related to single words and therefore if this change is missed, there will be more probabilities of Accuracy errors. In the case of MT, the engine might produce an output that is different in meaning, which would need to be completely rearranged or rewritten, thus causing Mistranslation errors. Overall, however, the No match category is not exempt from this type of error. Another interesting aspect is that Fuzzy match had a low percentage of Style errors, indicating the high quality of the original TM, and MT had a low percentage of Terminology errors. We are unsure about the reasons for this, since the same TM was used to train the engine. It could be that translators when correcting blatant errors in MT segments consulted the glossary more frequently.

But are these differences in number of errors significant? A Poisson regression model with repeated measures taking as a variable *Total errors* and an offset equal to the *text length* was set up and statistically significant differences were found between the three types of translations (F=52.48 and p<0.0001). There are differences between the three Match categories, but are results significantly different for MT and Fuzzy matches? From this model, the estimated means for the *logarithm of total errors /segment length* were calculated according to the match category. To better interpret the results, Table 9 presents the corresponding estimates expressed in the number of errors per segment length depending on the match category (length Fuzzy match=618 words, MT match = 757 words, No match = 749 words).

Match	Mean	SD	Lower	Upper
Fuzzy match	7.06	0.45	6.23	8.00
MT match	7.50	0.47	6.63	8.49
No match	12.35	0.70	11.03	13.82

Table 9: Estimated mean of errors

The estimated mean of error in Fuzzy match is 7.06 with a confidence interval of (6.23, 8), in MT match, 7.5 with a confidence interval of (6.63, 8.49), and finally in No match 12.35 with a confidence interval of (11.03, 13.82). It can clearly be seen that the No match is significantly different whereas the other two categories, Fuzzy and MT match, are not. Translators made more errors when translating without a proposal and made very similar amounts of errors when editing text from machine translation or translation memories fuzzy match segments from the 85-94 percent range. Therefore, the number of final errors would not be affected by using MT. These results are different to those found in the pilot project (Guerberof 2008). It is difficult to establish the reasons for this, since the two projects are quite different in terms of engine used, volume of words, translation memories used, the text itself, the number of translators, the translators themselves, the reviewers, the instructions received and the statistical analysis. We can only hypothesize that the results now are more accurate due to the fact that the volume of words to translate and number of translators were higher, not to mention the fact that the instructions sent out to the translators were written with the experience drawn from the pilot project. This time around, we felt the need to clearly indicate that the source text (the English text) must be read first and we gave a clearer description of the desired quality. And again, this may be a reflection of the quality of the original translation memories and glossary. With regards to other studies that explore post-editing and productivity, Tatsumi (2010) did not assess the final quality of the post-edited text, and De Almeida and O'Brien (2010) assess the changes made in the post-edited text but not the quality of the resulting post-edited text. Plitt and Masselot (2010) arrive at a similar conclusion in their study as they establish that the sentences with errors in the translation sample were higher than those in the post-edited sentences for German, French, Italian and Spanish. Fiederer and O'Brien (2009) find that ten raters judged the post-edited 30 sentences into German to be of higher clarity and accuracy, while the translations were judged to be of better style. García (2010), in a project from English into Chinese using Google Translator Toolkit (GTT), found no significant differences between translations using MT and others described as "entirely human", and although reviewers gave different ratings, the overall assessment was favorable to MT-seeded texts. There are a number of commercial pilots that establish no difference in quality, including IBM (Roukos et al. 2011) and Sybase (Bier and Herranz 2011). In a Translation Automation User Society Conference (TAUS 2011) a PayPal representative stated that "human quality was not good enough for PayPal" (Dove et al. 2011) and she explained in detail the reasons for this, most importantly because MT clarified the meaning with heavily tagged segments, MT did not miss trailing spaces and it performed better in consistency. In a similar study to this one, De Sutter and Depraetere (2012: 13) conclude that "productivity increases without jeopardizing final translation quality" and that the quality of the post-edited translation and human translation is similar. Our results are therefore not that dissonant in relation to what is being discussed in similar studies and at conferences about this same topic.

7 Discussion

We saw in the Productivity section that translators had different levels of productivity gain when using the MT matches, and that these different levels could also be seen in the Time savings column. For example, TR 23 had a 6 percent gain while TR 6 had a 60 percent gain (see Table 3). Prices are determined according to these time savings. In the localization industry, payments for the 85-94 percent range tend to be 60 percent of the full word rate (see De Palma et al. 2008), which assumes a 40% time saving for this Fuzzy match range. Table 3 shows that only TR3, TR6 and TR12 are above 40 percent, and the rest have values below this. Some of these values are close to 40 (for example, TR18 or TR21) but others are quite far from achieving these savings (for example, TR11, TR20 or TR23). The average time saving for Fuzzy matches for all translators is 32 percent and 37 percent for MT matches. It is important to remember that the time savings in this project are obtained with a high quality TM and MT output. By looking at this data, it might appear that the standard pricing in the industry is lower than the one that should be applied to this particular set of fuzzy matches, and as a consequence to the MT matches. However, in our project, the translators were not working with the original tool (SDL Trados 2007) where changes to be made in each fuzzy match are highlighted, nor they are aware of the origin of the proposal (MT or TM), which could influence their confidence in the proposal received. Therefore, it is possible that the time savings when using the standard tool are higher than those obtained with the tool in the experiment, but this is not certain.

As we saw in our previous project (Guerberof 2008), there is great disparity between translators when it comes to productivity, making it difficult to use standard measurements and hence to set a correct pricing that would compensate everyone's real effort. If we consider the 2,500 words per day standard industry metric and this volume is paid per word, a slower translator will get paid less per day and a faster translator will be paid more per day, but the amount paid for the full number of words (the job) is still the same. In this sense, we can say that the payment system is "fair" if the job is considered. However, if a percentage of the word edit or post-edit is paid, for example 50 percent, assuming an average productivity increase for all translators alike (for the same job), then a translator that has a lower productivity increase when using these proposals than when translating on his or her own might benefit less than a translator that has a higher productivity increase when using these same proposals, and therefore this originally fast translator might be paid less for the same job. The situation is similar with translation memories and machine translation and it might be difficult to find a satisfactory solution to determine a "fair" price. In most cases, the translators should really analyze if the compensation scheme applied for a particular project is beneficial for them according to the productivity they experience during this job or series of similar jobs. Moreover, they should also consider if the use of MT and TM segments might benefit the quality they deliver, as we have seen in this study.

We cannot find a simple solution for pricing MT, but we can see that a translator might complain for the payment receive for MT output in relation to their productivity and this payment might be "fair" for another translator in relation to their particular productivity. We agree that there is a need for standardization in pricing but as we have seen in this study, translators with the same set of segments have very different productivities.

8 Conclusions

We find in this experiment that the time invested in processing MT match segments is not statistically different from the time invested in processing Fuzzy match segments in the 85-94 percent range. Further, the results show statistical differences in speed between No match, on the one hand, and Fuzzy and MT matches on the other, indicating that the texts proposed by the TM and MT help translators increase their productivity.

We also sought to investigate adequate levels of payment for MT matches, and several conclusions can be drawn from this experiment. If the engine is trained with sufficient (high number of bilingual or parallel data) and cleaned translation memories and the automatic and/or human evaluation scores are high (in our case the BLEU score was 0.60 and the human evaluation 4.5 out of 5), translators show similar productivity in processing MT output as in processing 85-94 Fuzzy matches. The productivity gain and associated time savings vary considerably for each translator and this indicates that some translators benefit more than others from these translation proposals. The average time savings for these translators (32 percent for Fuzzy matches and 37 percent for MT matches) is lower than the average 40 percent assumed for this type of match (85-94 fuzzy matches) in the industry but this might be explained in some cases by the use of a different tool that does not highlight the changes in the Fuzzy match segments. We can conclude that the productivity of MT matches can be similar to that of 85-94 percent TM matches if the MT output is of high quality, tagging is not excessive, instructions are short and clear, terminology is either correct in the output (no terminological updates have occurred after training the engine) or few glossaries need to be consulted (as in this experiment), and the quality expected is not significantly different to the quality of the output.

Turning to quality, the number of aggregated errors from three reviewers show that the final quality of the post-edited MT segments is not statistically different from the final quality of the edited Fuzzy match segments and it is significantly higher than the translated No match segments. These results might be surprising to some, because unaided human translation (in this case, translations produced in the No Match category) is sometimes assumed to be higher, with respect to quality, than post-edited translation (in this case, translations produced from the MT matches). However, our results are really quite logical given the quality of the original material (translation memories) used to train the engine and from which the Fuzzy matches were obtained, and given that the No match segments were only translated and not revised by a third party. Our goal was to find out if MT had an impact on the final quality of the post-edited text. We can conclude that in this experiment both the MT and TM proposals had a positive impact on the quality since the translators had significantly more errors in the No match category, translating on their own, than in the MT and Fuzzy match categories.

Acknowledgments

The research project this article stems from was funded by the European Association of Machine Translation (EAMT) and the Centre for Global Intelligent Content (CNGL). I would also like to thank Crosslang, MicroStrategy and Logoscript (now Pactera) for their key contribution to the project.

I'm especially grateful to Dr. Anthony Pym and Dr. Sharon O'Brien for their continuous support on the completion of the doctoral thesis that this project is part of.

References

Asia Online (2012) JABA-Translations machine translation workshop demonstrates MT equivalent to 85% fuzzy match productivity. Language Studio Newsletter. http://www.asiaonline.net/EN/Resources/Newsletters/201203.htm. Accessed January 2014

Autodesk (2011) Translation and post-editing productivity. http://langtech.autodesk.com/productivity.html. Accessed January 2014

Bier K, Herranz M (2011) MT experience at Sybase. Localization World Conference, Barcelona. http://www.slideshare.net/manuelherranz/loc-world2011-kbiermherranz-8730502. Accessed January 2014

Carl M, Dragsted B, Elming J, Hardt D, Jakobsen A (2011) The process of post-editing: a pilot study. In: Proceedings of the 8th International NLPSC Workshop. Copenhagen Studies in Language 41, Frederiksberg, pp 131-142 http://www.mt-archive.info/NLPCS-2011-Carl-1.pdf. Accessed January 2014

De Almeida G, O'Brien S (2010) Analysing post-editing performance: correlations with years of translation experience. In: Proceedings of the 14th EAMT Conference, St. Raphael. http://www.mt-archive.info/EAMT-2010-Almeida.pdf. Accessed January 2014

De Palma D, Sargent B, Bassetti T, Beninatto R (2008) The price of translation: a comprehensive analysis of pricing for globalization service buyers. Common Sense Advisory. www.commonsenseadvisory.com

De Sutter N, Depraetere I (2012) Post-edited translation quality, edit distance and fluency scores: report on a case study. In: Journée d'études Traduction et qualité Méthodologies en matière d'assurance qualité, Universtité Lille 3, Sciences humaines et socials, Lille. http://stl.recherche.univ-lille3.fr/colloques/20112012/DeSutter&Depraetere_2012_02_03.pdf. Accessed January 2014

Dove C, Way A, Johnson D (2011) Quality above price and speed. TAUS Conference, Santa Clara. http://www.youtube.com/watch?v=eZDAD7Y_MHE Accessed January 2014

Fiederer R, O'Brien S (2009) Quality and machine translation: a realistic objective? The Journal of Specialised Translation (11). http://www.jostrans.org/issue11/art_fiederer_obrien.pdf. Accessed January 2014

Flournoy R, Duran C (2009) Machine translation and document localization at Adobe: from pilot to production. In: Proceedings of the 12th MT Summit, Ottawa. http://www.mt-archive.info/MTS-2009-Flournoy.pdf. Accessed January 2014

García I (2010) Is machine translation ready yet? Target 22(1):7-21, Amsterdam and Philadelphia, Benjamins

García I (2011) Translating by post-editing: Is it the way forward? Machine Translation 25(3):217-237, Netherlands, Springer

Guerberof A (2008) Productivity and quality in machine translation and translation memory outputs. Dissertation, Universitat Rovira i Virgili. http://bit.ly/1a83G9p. Accessed January 2014

Guerberof A (2012) Productivity and quality in the post-editing of outputs from translation memories and machine translation. PhD Thesis, Universitat Rovira i Virgili. http://www.tdx.cat/handle/10803/90247. Accessed January 2014

Koehn P, Hoang H, Birch A, Callison-Burch C, Federico M, Bertoldi N, Cowan B, Shen W, Moran C, Zens R, Dyer C J, Bojar O, Constantin A, and Herbst E (2007) Moses: Open source toolkit for statistical machine translation. In: Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, ACL, Stroudsburg, pp 177–180

O'Brien S (2011) Towards predicting post-editing productivity. Machine Translation 25(3):197-215, Netherlands, Springer

O'Brien S (2002) Teaching post-editing: a proposal for course content. In: Proceedings for the 6th EAMT Conference, Manchester, pp 99-106. http://mt-archive.info/EAMT-2002-OBrien.pdf. Accessed January 2014

O'Brien S (2006a) Methodologies for measuring correlations between post-editing effort and machine translatability. Machine Translation 19(1):37-58, Netherlands, Springer

O'Brien S (2006b) Eye tracking and translation memory matches. Perspectives: Studies in Translatology 14 (3): 185-205

Offersgaard L, Povlsen C, Almsten L, Maegaard B (2008) Domain specific MT use. In: Proceedings of 12th EAMT Conference, Hamburg, pp 150-159. http://www.mt-archive.info/EAMT-2008-Offersgaard.pdf. Accessed January 2014

Papineni K, Roukos S, Ward T, Zhu W.J. (2002) BLEU: A method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the ACL, Stroudsburg, pp 311-318. http://acl.ldc.upenn.edu/P/P02/P02-1040.pdf. Accessed January 2014

Plitt M, Masselot F (2010) A productivity test of statistical machine translation post-editing in a typical localisation context. The Prague Bulletin of Mathematical Linguistics, Prague, pp 7-16. http://ufal.mff.cuni.cz/pbml/93/art-plitt-masselot.pdf . Accessed January 2014

Roukos S, Rojas F, Ming Xu J, Pont Nesta S, Martínez Corriá A, Chapman H, Vohra S (2011)The value of post-editing: IBM case study. Localization World Conference, Barcelona. www.localizationworld.com/lwbar2011/presentations/files/E6.ppt Accessed January 2014

Tatsumi M (2010) Post-editing machine translated text in a commercial setting: observation and statistical analysis. PhD Thesis, Dublin City University. http://doras.dcu.ie/16062/. Accessed March 2012