Productivity and quality in the post-editing of outputs from translation memories and machine translation

Ana Guerberof Arenas
PhD programme in Translation and Intercultural Studies
Universitat Rovira i Virgili, Tarragona, Spain
Ana.Guerberof@gmail.com

Abstract

Machine-translated segments are increasingly included as fuzzy matches within the translation-memory systems in the localisation workflow. This study presents preliminary results on the correlation between these two types of segments in terms of productivity and final quality. In order to test these variables, we set up an experiment with a group of eight professional translators using an on-line post-editing tool and a statistical-based machine translation engine. The translators were asked to translate new, machine-translated and translation-memory segments from the 80-90 percent value range using a post-editing tool without actually knowing the origin of each segment, and to complete a questionnaire. The findings suggest that translators have higher productivity and quality when using machine-translated output than when processing fuzzy matches from translation memories. Furthermore, translators' technical experience seems to have an impact on productivity but not on quality.

Keywords: *Translation memory, machine translation, post-editing, revision, productivity, quality, errors, editing, professional translators, experience, fuzzy match, processing speed, localisation*

Introduction

New technologies are creating new translation processes in the localisation industry, as well as changing the way in which translation is paid for. In the past, translation involved precisely that, the translation of entire software, documentation and help materials into new target texts for the local markets. As localisation matured, translation memories (TM) were created and texts were recycled in different but rather similar projects. Productivity increased and consequently prices of translations decreased. Since the 1980s, machine translation (MT) technology has improved significantly and has been incorporated into the localisation workflow as another type of translation aid, rather than attempting to have a fully automatic high-quality translation. It remains to be seen what effect this technological development will have on pricing structures.

Major software development companies now pre-translate source text using existing translation memories and then automatically translate the remaining text using a machine-translation engine. This "hybrid" pre-translated text is then given to translators to post-edit. Following guidelines, the translators correct the output from translation memories and machine translation to produce different levels of quality. Gradually this activity, post-editing, is becoming a more frequent activity in localisation, as opposed to the full translation of new texts.

In an industry that moves so rapidly, there is more focus on finalising projects than on the process itself. Therefore these translation aids are used in the localisation workflow with limited data to quantify the actual translation effort and the resulting quality after post-editing. Since productivity and quality have a direct impact on pricing, it is of capital importance to explore that relationship in terms of productivity and quality of the post-editing of texts, coming from translation-memory systems and machine-translated outputs, in relation to translating texts without any aid.

In this context, it seems logical to think that if prices, quality and times are already established for TMs according to different level of fuzzy matches then we only need to compare MT segments with TM segments, rather than comparing MT output to human translation. Therefore, once the correlation is established, the same set of standards for time, quality and price can be used for the two types of translation aid.

Preliminary premises

After a study by Sharon O'Brien (2006) where she establishes a correlation between MT segments and TM segments from the 80-90 percent category of

fuzzy match, we formulated our initial hypothesis. This one was that *the time invested in post-editing one string of machine translated text will correspond to the same time invested in editing a fuzzy matched string located in the 80-90 percent range.* This hypothesis is predicted on the assumption that the raw MT output is of reasonable quality according to the Bleu Score (Papineni et al 2002, p. 311).

Measuring productivity on its own, as in our first hypothesis does not make sense if it is not done in relation to an equal level of final quality. If the time necessary to review MT segments is greater than the time necessary to review New or TM segments, the productivity gain made during the translation and post-editing phase would be offset by the review phase. Therefore, we claimed that *the final quality of the target segments translated using MT is not different to the final quality of New or TM segments.*

Localisation has a very strong technical component because of the content as well as the tools required. On many occasions we associate technical competence with speed, that is, the more tools we use the more automated the process becomes and the less time we spend completing a project. Therefore, our third hypothesis claimed *that the greater the technical experience of the translator, the greater the productivity in post-editing MT and TM segments.*

Methodology

In order to prove our hypotheses we carried out an experiment with nine subjects. One subject carried out the preliminary test and the remaining eight performed the actual pilot experiment. The translators used a web-based post-editing tool to post-edit and translate a text from English into Spanish. The text had 791 words; 265 words of new segments (new text to translate), 264 words of translation-memory segments (Trados was used to create the fuzzy matches) and 262 words of machine-translated segments (Language Weaver's statistical-base engine was used to create the output). We selected a supply-chain software product for the corpus as we wanted to use typical content from the localisation industry. At the end of their assignment, the subjects filled in a questionnaire with information related to the pilot experiment and their own experience in the field. The final output was then revised, errors were counted and conclusions drawn.

Experiment design

Translators

We contacted a group of nine professional translators, five women and four men, with ages ranging from 22 to 46 years. They all have first degrees or Master's Degrees in Translation. Their experience ranges from 1 year to more than 10 years in the translation industry and most have specific experience in localisation. They were contacted by email in all cases and they received no training to carry out the pilot experiment, only a set of instructions. The translators were not paid for the work that they carried out and although they knew the work was for research, and they might have inferred from the tool that the research dealt with machine translation, they were not given any specific information on the topic. Due to the fact that they were professional translators working for a short period of time and that they knew their work would be part of a research project, we would assume they maintained their usual working standards.

Training the engine

We provided Language Weaver with a translation memory containing 1.1 million words and a core glossary. They then created a customized engine using the relevant translation memories and a validated terminology list. Finally, they uploaded these segments into the post-editing tool.

Creating the translation memory segments

For our research we needed to create a file containing segments in the 80-90 percent category to feed these lower fuzzy matches into the tool. To prepare the file, we pre-translated existing html files from a help project of the supply-chain software with a previous memory in order to obtain fuzzy matches using the option Pre-translate in SDL Trados (version 7.1). We created ttxs files with different fuzzy match values. We then exported all segment pairs together with their corresponding fuzzy level (54, 75, 86 and so on) to Excel. This was done with a small tool created specifically for this purpose called Slicer.

Since we only needed a small number of words and not all of the segments, we randomly selected a number of segments from each category using the function *Random.between* in Excel. This gave us the desired number of segments in a random selection.

Post-editing tool

The translators were able to connect to the post-editing tool online. They could then translate/post-edit

the proposed segments of text without knowing their origin (MT, TM or New segments) and the tool measured the time taken in seconds for each task. The post-editing tool required the translator to log on with a specific user name and password, so each translator could only see the text assigned to them. Once they opened the task, they were presented with a screen containing the actual task as seen in Figure 1.

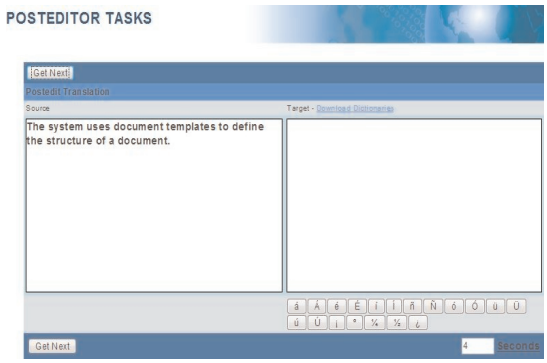

Figure 1: Web-based tool for post-editing tasks

The Source window contained the source text in English, and the Target window contained either a blank screen or a proposed text in Spanish. The Spanish text was either a MT or TM segment. Once finished with a string, the translator had to click on the Get Next button and proceed with the following segment until they had reached the end of the assignment.

Questionnaire
The aim of the questionnaire was to define the translators' experience in localisation, tools, subject matter and post-editing MT. The questionnaire consisted of 17 questions that addressed these areas. It contained dichotomous questions, questions based on level of measurement and filter questions. The main aim of the questionnaire was to describe the group of translators and establish their experience in localisation, supply chain, knowledge of tools, and post-editing MT, as well as gather their views on MT. We matched the answers from the translators to the processing speed from the tool and the number of errors in the final sample.

Quality of the samples
The final target texts were checked to see the number of final errors in each sample. This could give us an indication of productivity versus quality. If translating with an aid was faster than the human translation, but there were more errors, then more time would be needed in a final review phase, thus altering the real translators' productivity.

We used LISA standards to measure and classify the number of errors. We classified the errors according to their source (New, MT or TM segments) to see if each category had similar number of errors. We classified errors according to type to see their frequency in each type of segment. Lastly, we matched the errors with the productivity by means of a coefficient of error based on the average revised word per minute.

Results

Productivity

Processing speed
Processing speed is the processing time in relation to the words processed in that time, that is, words divided by time. The number of words was almost identical in the three categories, New (265 words), MT (262 words) and TM (264 words) consequently our processing times and processing speeds were not notably different. The results are given in Table 1. We have highlighted in bold the maximum and minimum values per segment category.

Translator	New	MT	TM
TR 1	12.12	18.69	14.52
TR 2	10.76	10.28	10.75
TR 3	**22.08**	**21.21**	16.40
TR 4	8.55	9.79	10.22
TR 5	**5.85**	12.04	8.18
TR 6	8.11	9.12	**8.08**
TR 7	20.03	20.77	**18.48**
TR 8	7.42	**8.96**	10.47

Table 1: Translators' processing speed in words per minute per segment category

This table shows that four out of eight translators performed faster using MT (TR 1, TR 5, TR 6, and TR 7), two were faster translating New segments (TR 2 and TR 3), and two were faster processing TM segments (TR 4 and TR 8). In total, six were faster using a translation aid than translating without any aid. Only TR 2 shows the slowest processing speed when using MT by quite a small margin in comparison to New or TM segments.

Let us have a look at the statistical summary:

Translator	New	MT	TM
Mean	**11.87**	**13.86**	**12.14**
Median	**9.66**	**11.16**	**10.61**
Std. Deviation	**6.02**	**5.40**	**3.87**
Max	22.08	21.21	18.48
Min	5.85	8.96	8.08
Range	16.23	12.25	10.41
1st Quartile	7.94	9.62	9.71
3rd Quartile	14.10	19.21	14.99
Diff quartiles	6.16	9.59	5.28

Table 2: Statistical summary of processing speed

Table 2 shows, in bold, that translators process, on average, more words per minute in MT than in TM or New segments and that they process, in turn, more words in TM than in New segments. All the same, the standard deviation is extremely high, 6.02 for New segments, 5.4 for MT and 3.87 for TM. For example, the range of variation (seventh row) between the maximum and minimum values is 16.23 words in New segments, 12.25 in MT segments and 10.41 in TM segments. Hence the mean, as a unique value is, not a fully representative number for the data shown here. The median for all the values, in bold, tells us that MT continues to be faster than human translation (approximately 16 percent) and faster than using TM (approximately 5 percent). The first quartile (eighth row) shows that processing TM segments is faster than processing New or MT segments, only 1 percent higher than MT, and in turn MT is faster than processing New segments, by approximately 21 percent. In this case, the quartile analysis shows that the translators that process fewer words per minute have a higher correlation between TM and MT than the group that processes more words. The second quartile, equivalent to the median, shows that MT is faster than New and TM segments, although the difference between MT and TM values is not very pronounced. In the third quartile, ninth row, we see that the speed for New segments and TM is extremely close, while MT is definitely faster. The difference between the first and third quartile, tenth row, shows us that there are pronounced differences, especially in MT with 9.59 words difference, then in New with 6.16 and in TM with 5.28 words.

Productivity gain

The productivity gain is the relationship between the number of words translated per minute per single translator without any aid and the number of words translated per minute by the same translator with the aid of a tool, TM or MT. This value is expressed as a percentage value.

In Table 3 we see the statistical summary regarding productivity gain:

Translator	MT vs. New	TM vs. New
Mean	**25%**	**11%**
Median	**13%**	**10%**
Std. Deviation	37%	23%
Max	106%	41%
Min	-4%	-26%
Range	110%	67%
1st Quartile	2%	-2%
3rd Quartile	29%	25%
Diff quartiles	27%	27%

Table 3: Statistical summary of productivity gain

The mean values in MT and TM in relation to New segments show us that translators have a higher productivity gain if they use a translation aid. The gain was higher in MT segments than in TM segments, with 25 and 11 percent respectively. Nonetheless, the standard deviation is extremely high and the range of variation is very pronounced. The median value, in bold, shows that MT has a higher productivity gain (13 percent) but that the difference with TM is not very pronounced (10 percent). In the first quartile, eighth row, the productivity gain provided by the translation aid, MT or TM, is not very pronounced, and relatively similar (4 percent variance). Still the productivity gain for TM is negative, indicating a decrease in productivity. This quartile includes TR 2, TR 3, TR 7 and TR 6. In the third quartile, the productivity gain for both MT and TM is higher (29 and 25 percent respectively). This quartile includes TR 4, TR 5, TR 8 and TR 1. The highest productivity gain, if we take the statistical values, never goes over 29 percent (third quartile using MT). We should remark that the values in the quartiles correspond partly to the faster and slower translators and this seems to indicate that faster translators take less advantage of translation aids than do slower translators.

Quality

Existing errors and changes in MT and TM

Before we looked at the errors found after the assignment was completed, we needed to look at the number of errors and corrections existing in the MT and TM segments before the pilot took place. Otherwise,

if we found that one category, MT or TM, contained more errors than the other, it would have been logical, although not necessarily true, to assume that there would be more errors after the assignment was completed in that same category. Similarly, we classified the errors found using the LISA standard and we had identified the number of changes that were necessary to perform in the TM segments.

The TM segments contained 1 Mistranslation, 1 Accuracy, 1 Terminology and 2 Language errors. These five errors came from the legacy material used to build the translation memory and were therefore made by human translators. There were 17 changes needed in the text. These changes were text modifications, insertions and deletions between the original source text and the new source text. This meant that there were 5 existing errors and 17 changes to make in the TM segments.

On the other hand, the MT segments contained 25 Language and 2 Terminology errors, a total of 27 existing errors in the MT segments. The typical errors found in MT output were wrong word order, grammar mistakes (concordance of verb and subject, concordance of genre) and inconsistent use of upper and lower cases. There were also a couple of cases where the MT engine chose the wrong term for the cotext given.

A priori, the number of existing errors and changes in TM versus the ones in the MT segments was very similar: 22 in the TM segments versus 27 in the MT segments, and this meant that the source text should not necessarily condition the final target text. The actual process needed to correct the texts was different in our view. This was due to the fact that the TM segments, on the one hand, needed insertions, changes and deletions where it was necessary to constantly refer to the source text, as well as 5 "standard" errors where the main reference was the target text. On the other hand, MT errors involved mainly language changes that were quite distinct and where a constant reference to the target text was necessary because they involved changing the word order, use of verb tenses, use of upper and lower cases and concordance of number. This difference in the required post-edit approach could mean different results in the final text depending on where the focus was when translators were working on the target text. It is important to mention at this point that translators did not know the origin of the segments (MT or TM) and obviously if these segments were full (100 percent) or fuzzy matches (54-99 percent).

Error analysis

We used the LISA form in the eight samples and we counted the errors according to its classification and according to the type of segment in order to compare the results. The classification of errors was carried out by the researcher mainly due to time and budget limitations and also because the researcher had extensive experience in reviewing these type of texts in this language combination. The texts were corrected and then compared against each other to assure that the same classification criteria were followed in all texts.

Table 4 shows the final number of errors per translator according to the type of segment, and the total number of errors. The table is sorted according to ascending total errors. Totals are highlighted in bold.

Translator	New	MT	TM	Totals
TR 3	1	1	4	**6**
TR 2	2	3	6	**11**
TR 4	2	5	6	**13**
TR 1	2	3	10	**15**
TR 6	4	5	8	**17**
TR 8	6	3	9	**18**
TR 7	7	5	9	**21**
TR 5	3	9	13	**25**
Totals	27	34	65	**126**

Table 4: Number of errors per type of segment and translator

Table 4 shows that all segment categories contain errors, and all translators have errors in all categories. There are a total of 126 errors in the final texts. A total of 27 errors are found in the New segments and 99 in the combination of TM and MT segments. Translators did not have the possibility, when using the tool, to go back and correct their own work and the segments have not been reviewed by a third party. We nevertheless see that in all eight cases there are more errors in TM segments than in any other category. In five out of eight cases, there are more errors in MT than in New segments (TR 1, TR 2, TR 4, TR 5 and TR 6); in two cases (TR 7 and TR 8) there are more errors in New than in MT segments; and in one case there is an equal number of errors in both New and MT (TR 3).

The first striking result is that the number of errors in TM segments (65) is 141 percent higher than that of the New segments (27) and 91 percent higher than that of the MT segments (34). MT segments, on the other hand, contain 26 percent more errors than New

segments. We find that the number of errors in TM segments is consistently higher in all eight cases while the errors for New and MT segments vary among the subjects.

Errors per type

We have analysed how errors are distributed according to the LISA standard to see if the typology of errors varies depending on the type of source text, in order to understand if the type of text has an effect on the number of errors. We can see this analysis in Table 5:

Type of error	New	MT	TM	Totals	% New	% MT	% TM	% Total
Mistranslation	10	2	8	20	**8%**	**2%**	**6%**	**16%**
Accuracy	9	14	34	57	**6%**	**11%**	**27%**	**44%**
Terminology	2	9	9	20	**2%**	**7%**	**7%**	**16%**
Language	6	8	14	28	**6%**	**6%**	**11%**	**23%**
Consistency		1		1	**0%**	**1%**	**0%**	**1%**
Totals	**27**	**34**	**65**	**126**	**21%**	**27%**	**52%**	**100%**

Table 5: Number and percentage of errors per type of error

There are 57 Accuracy errors that represent 44 percent of the total number of errors (almost half of the errors), and 34 of them, that is 27 percent of all the errors, are found in the TM segments. There are 9 Accuracy errors in New segments and 14 in MT, representing 6 and 11 percent respectively. One possible explanation for this number of errors in the TM segments could be that when translators are presented with a text that flows "naturally" like a human translation they seem to pay less attention to how accurate that sentence is. On the other hand, because errors in MT segments are so obviously wrong, the mistakes seem to be easier to detect. As we explained above, most of the changes in TM required the translator to look at the source text and not just focus on the proposed target. The fact that the TM segments have so many errors could be explained by the fact that translators possibly consulted the source text less than they would have if they had been translating a new text with no aid. We have seen in previous studies that monolingual revision is less efficient than bilingual revision (Brunette et al. 2005), that there is a trend towards error propagation in the use of TMs (Ribas 2007), and that using TMs increased productivity, but "translators using TMs may not be critical enough of the proposals offered by the system" (Bowker 2005, p.138) and they left many errors unchanged.

In our study there are 29 Language errors that represent 23 percent of the total number of errors: 14 of them, that is 11 percent, are found in TM segments while 6 and 8 (6 percent) are found in New and MT segments respectively. We see again in this case that the TM contains the most errors and, again, this could be due to the reasons explained above: when translators are provided with a text that flows naturally they seem to accept the segments as they are without questioning the text correctness. It is true that some errors could have been spotted on a second review, but we can say that errors in TM were not as frequently spotted as the ones in the MT segments.

From the 20 mistranslation errors, 10 are found in the New segments, representing 8 percent of the total, 8 errors are found in TM and only 2 mistranslation errors are found in MT representing 6 and 2 percent respectively. The fact that there are so few mistranslation errors in MT segments might indicate that using MT helps translators clarify possibly difficult aspects of the source texts thus improving general comprehension of the text.

From the 20 Terminology errors, only 2 are found in the New segments as opposed to 9 in both MT and TM segments. This seems to indicate that translators tend to consult the existing glossaries more when they are presented with new texts, rather than questioning the proposed terminology used in MT and TM. It might be logical not to check terminology in a pre-translated text, but terminology is not always correct in TM and MT outputs due to updates and changes in existing terminology. This indicates that instructions should be provided to reviewers or translators to specifically check glossaries or, alternatively, terminological changes need to be made directly to the TM or MT before the translation process begins.

The consistency error found in the MT segments that represent 1 percent of the total is related to the inconsistent use of upper and lower cases and it is a reflection of a known issue in MT output. We would venture that if the translators had received specific instructions on output error typology, this error would have been corrected.

Errors vs. productivity

We have established that an increase in productivity cannot be considered in isolation from the quality of the samples. So how does the number of errors found in the samples affect the overall productivity of the translators? Can we say that using MT or TM decreases or increases the productivity of a translator taking into account the final errors? To find an answer to these questions, we decided to penalise translators in their processing speed according to the number of errors made. To do this, we calculated a general coefficient of error to be used as a form of penalty (or correction) in words per minute and then we applied this coefficient to the processing speed of the eight subjects in order to see the impact of errors on the productivity gain.

Calculation of the error coefficient

We realised that the best way to determine the error coefficient would be to measure the reviewing time of these segments in a standard revision process by a third party. In this case, because the review is not part of the scope of this study, we took the metrics used for reviewers of localisation texts; approximately 7500 words per day (this figure may be higher or lower depending on the metric used by each individual localisation agency). With this figure in mind, we established that a reviewer reviews 0.26 words per minute (if we took a higher figure the value would be of course higher). We took the number of errors per translator and we applied the coefficient of error for each source of error and then recalculated their processing speeds, thus obtaining a final figure that reflected the impact of errors on their processing speed.

Once we had the new processing speeds for all translators, we recalculated the productivity gain comparing the different categories in order to see the impact on productivity that the errors might have had in a working environment. Negative values are highlighted in bold.

Translator	Total processing speed	MT vs. New	TM vs. New
TR 1	41.43	54%	3%
TR 2	28.93	**-7%**	**-10%**
TR 3	58.53	**-4%**	**-30%**
TR 4	25.18	6%	8%
TR 5	19.57	91%	**-5%**
TR 6	20.89	11%	**-15%**
TR 7	53.82	7%	**-11%**
TR 8	22.17	40%	39%

Table 6: Translators' productivity gain minus coefficient of error

In Table 6, MT is still faster than translating with no aid in six out of eight subjects (TR 1, TR 4, TR 5, TR 6, TR 7 and TR 8). The other two subjects (TR 2 and TR 3) have a negative value. This value has increased for TR 2 and remained stable for TR 3 (who made an equal number of errors in MT and TM categories), and in both cases the negative value is never below 7 percent. TR 4, TR 6 and TR 7 show a positive value of around 10 percent. On the other hand, TR 1, TR 5 and TR 8 show a positive value above 40 percent. Even if errors are considered, using MT is still more productive than no aid at all.

If we look now at the productivity gain of TM, the changes are more pronounced. Five out of eight cases have a negative productivity when compared to New segments (TR 2, TR 3, TR 5, TR 6 and TR 7), and in four cases the negative value is equal to or below minus 10 percent. In the case of TR 3, the value goes down to minus 30 percent. In two other cases (TR 1 and TR 4) TM brings a slight productivity increase with 3 and 8 percent respectively. Only the remaining case (TR 8) seems to have a pronounced productivity increase, with 39 percent. If errors are considered, using TM fuzzy matches (80-90 percent) does not appear to be productive when compared with translating without any aid.

In brief, if we consider errors when calculating the productivity gain, we see that although MT seems to play an important role in increasing productivity in most cases, TM has the opposite effect. It is important to remark here that we are referring to segments that belong to the 80-90 percent category of fuzzy match and not TM segments that include all levels of matches. It could well be that this translation memory as a whole provides a productivity increase for translators. But the 80 to 90 percent category of fuzzy matches does not appear to do so, and this is remarkable if we consider that these segments tend to be paid at 60 percent of their value (the global price including review), thus assuming a 40 percent productivity gain, and that this productivity was not achieved by any of our translators when errors are considered.

Table 7 shows the statistical summary of the new productivity gain. Mean and Median values are highlighted in bold.

Translator	MT vs. New	TM vs. New
Mean	**25%**	**-3%**
Median	**9%**	**-8%**
Std. Deviation	34%	20%
Max	91%	39%
Min	-7%	-30%
Range	98%	68%
1st Quartile	3%	-12%
3rd Quartile	43%	4%
Diff quartiles	40%	16%

TABLE 7: STATISTICAL SUMMARY OF PRODUCTIVITY GAIN MINUS COEFFICIENT OF ERROR

The correlation between MT and TM in relation to New segments shows that translators have a higher productivity gain if they use MT but a negative productivity gain if they use TM (80-90 percent matches). The range of variation is very pronounced (TR 5 has a value of 91 percent as opposed to TR 2 who has -7 percent). If we take the mean values, in bold, we see that MT has a productivity gain of 25 percent while TM presents a negative value of minus 3 percent in comparison to the previous positive value of 11 percent. The median values for both MT and TM have changed from 13 to 9 percent in MT and from 10 to minus 8 percent in TM. The first quartile shows that the productivity gain provided by MT is small with just 3 percent and negative in the TM with minus 12 percent. In the third quartile, the productivity gain for both MT and TM is positive (43 and 4 percent respectively).

Technical experience

Our third hypothesis claimed that the greater the technical experience of the translator, the greater the productivity in post-editing MT and TM segments. The first question that comes to mind is "What does technical experience mean?" We are aware that the term embraces several aspects of a translator's competence. For the purpose of this study we have defined technical experience as a combination of experience in localisation, in knowledge of tools, in subject matter (in this case supply chain), and in post-editing of machine translated output.

We obtained this data from the questionnaire that was provided to the translators at the end of the assignment. This data was then contrasted with the translators processing speed and number of errors to see if there was a correlation between technical experience, processing speed and errors. We took the processing speed as a result of the experiment without including the coefficient of error because we analyzed the errors separately. We took the mean in the processing speed as the number of subjects was smaller than in the productivity section, in the sense that all subjects were grouped according to experience thus decreasing the number of subjects per group, and the mean and median obtained were in most cases the same value.

The fact that the group was small and that the data obtained in terms of processing speed was dispersed made drawing final and general conclusions on any correlation between technical experience and productivity difficult. Nevertheless, we think it was necessary to correlate the processing speed obtained from the post-editing tool, errors and the questionnaire, even if it served only to test our methodology.

Summary data on translators' experience

To summarize: data that includes experience in localisation, knowledge of tools, supply chain and post-editing, we singled out the translators that showed more experience in all of the above sections. The translators that declared having more experience in the four areas were TR 3, TR 4, TR 5 and TR 7. The translators with less experience were TR 1, TR 2, TR 6 and TR 8. We took the mean value for each group of translators in relation to the processing speed and number of errors. Table 8 shows these results:

	Processing speed			Number of errors		
Experience	**New**	**MT**	**TM**	**New**	**MT**	**TM**
More	14.13	15.95	13.32	3.25	5.00	8.00
Less	9.60	11.76	10.95	3.50	3.50	8.25

TABLE 8: OVERALL EXPERIENCE VS. PROCESSING SPEED AND NUMBER OF ERRORS

The table shows that experience has a clear effect on the processing speed. The experienced group is faster than the group with less experience. We can see that the faster group is faster when working with MT than with New segments and TM (in this order). The slower group is also faster when working with MT segments than with TM and finally with New segments. The translators with less experience seem to make better use of both translation aids than the ones with more experience. Additionally, we see that the translators with no experience have very similar processing speeds for MT and TM segments (as we claimed in our first hypothesis).

The total number of errors is slightly higher in the experienced group than in the one with little experience, by 1 error. The number of errors in MT is high-

er in the experienced group by a small margin, 1.5 errors when compared to New and TM segments. This could be due to the fact that translators with more experience are more accustomed to MT output and this familiarity prevents them from seeing very visible errors precisely due to this familiarisation.

Final conclusions

Conclusions on productivity
Considering the mean value, the processing speed for post-editing MT segments is higher than that for TM and New segments. And post-editing TM segments, in turn, is faster than translating New segments. The data dispersion is nevertheless quite pronounced, with very high standard deviations and great differences between maximum and minimum values. The standard deviation is higher for processing New segments than for processing MT or TM segments which might indicate that using pre-translated segments slightly standardizes processing speed.

The fastest overall processing time results from translating New segments without any aid, while the translator with slowest processing time took advantage of MT and TM. This low productivity is more pronounced for TM than for MT. If we look at the productivity gains, the translators with lower processing speeds seem to take more advantage of the translation aids than the translators with higher processing speeds. We would need further research to confirm this trend.

The productivity gain, when compared to New segments, for translation aids is between 13 and 25 percent for MT segments, which is higher than the percentage reported by Krings (2001) and lower than the figures reported by Allen (2005) and Guerra (2003), and from 10 to 18 percent for TM segments.
Our first hypothesis is thus not validated in our experiment since MT processing speed appears to be higher when compared to the processing speed in TM fuzzy matches. The correlation between MT and TM is quite close in the groups that processed fewer words per minute. There exists, however, a pronounced difference in the groups that processed more words per minute, where MT ranks higher. The deviation is high, nevertheless, and we cannot draw concrete conclusions as productivity seems to be subject dependant. Krings (2001) also found that in measuring processing speeds, the variance ranged from 1.55 to 8.67 words per minute. Although O'Brien (2006) offers an average processing speed across four subjects without mentioning any deviation values she highlights (2007) that there can be significant individual differences in post-editing processing speed in-line with these findings.

Conclusions on quality
Overall we can say that there are errors in all translators' texts and errors are present in all three categories: New, MT and TM. This seems to be logical, considering that the tool did not allow the translators to go back and revise their work, and that no revision work was done afterwards by a third party.

More than half the amount of total errors, 52 percent, can be found in the TM segments, 27 percent in MT segments and 21 percent in New segments. The high number of errors in TM could be explained by the fact that the text flows more "naturally" and translators do not go back and check the source text, they just focus on the target text, while the MT errors are rather obvious and easier to spot without having to check the source text.

The number of errors in TM is higher than in any other category for all translators. On the other hand, the number of errors in MT is greater than in New segments in five out of eight cases. In two cases, there are more errors in the New than in the MT segments and in one case there is equal number of errors.

Accuracy errors represent the highest number of errors, 44 percent, and they represent the highest value in TM and MT. This seems to indicate that translators do not question the TM or MT proposal and do not check the source text sufficiently to avoid this type of error. Mistranslation errors had the highest value in New segments, but it is very low in MT segments. This could indicate that MT clarifies difficult aspects of the source texts, although more data is needed to explore this trend. Terminology errors are lower in New than in MT and TM segments, indicating that translators tend to accept the proposed terminology in MT and TM without necessarily checking the terms in the glossaries. This might lead to a recommendation that terminological changes or updates be made before starting the translation process or that the translators be instructed to check the glossary often.

The four fastest translators account for 53 errors while the four slowest translators account for 73 errors, which might indicate that the fastest translators tend to make fewer errors and vice-versa, although this is not true for all cases. The reason behind this difference could be that some translators

found the assignment more difficult than others, but at any rate this difference does not indicate an improved quality.

When a coefficient of error is applied, based on an average review speed per minute, to the processing speed, productivity decreases for all segments and in particular for TM segments. This is only applicable to matches from the 80-90 percent category. MT, on the other hand, presents a productivity increase in relation to translating New segments. The increase is higher than 7 percent as was presented in Krings' study (2001), and it seems to be located between 9 and 25 percent. Krings finds that when comparing existing errors in the output with actual errors found after post-editing, the translators are rated at 3.38 (in a range from 1 to 5) covering almost 80 percent of all the errors in MT. In our case the difference in errors between New and MT segments is not very pronounced, but the errors are quite high in TM segments. As far as we know, other research such as O'Brien (2006), Guerra (2003) and Allen (2003 and 2005) does not offer a matrix of final errors and consequently we do not really know how increases in productivity related to the final quality of their samples. O'Brien (2007) mentions the issue of quality and promises to address the topic in a follow-up study. The forthcoming article will be published in the Journal of Specialised Translation (2009).

The pilot study thus indicates that using a TM with 80 to 90 fuzzy matches produces more errors than using MT segments or human translation. The reason behind this could be that translators trust the content that flows naturally without necessarily critically checking accuracy against the source text.

Finally, our second hypothesis is not proven true by the pilot study as our results show that the quality produced by the translators is notably different when they use no aid, MT or TM, although the number of errors found in MT segments is closer to those found in New segments.

Conclusions on translators' experience

If we consider the results obtained we can say that experience has an incidence on the processing speed. Translators with experience perform faster if the average is considered. Similar to the findings by Dragsted (2004) when comparing the processing speed between students and professionals, translators with less experience in our pilot are slower than the ones with more experience.

The data on errors is not conclusive, as the difference between experienced and less experienced translators is none or very small. In the summary data on translators' experience, experienced translators have a higher number of errors in MT and in New segments when compared to the group with less experience. This could be explained by the small number of subjects, or the possibility that translators with more experience grow accustomed to MT type of errors and they do not detect them as easily as a "newcomer" to the field. The translators with less experience have more errors in TM but less in MT and New.

We could say that our third hypothesis is partially proven because translators with greater technical experience do have higher processing speeds in both MT and TM overall. It is important to point out as well that experience does not seem to have an impact on the total number of errors.

There is a strong need to further explore how new technologies are shaping translation processes and how these technologies are affecting productivity, quality and hence pricing. If translators and the translation community as a whole acquire more knowledge about the actual benefits of the tools in real terms, we can be prepared to come into the negotiating arena with the knowledge necessary to reach common ground with translation buyers.

References

Allen, J. 2003. "Post-editing". In Computers and Translation: A Translator's Guide. Harold Somers, ed. Amsterdam & Philadelphia: Benjamins. pp. 297-317.

Allen, J. (2005). "An introduction to using MT software". The Guide from Multilingual Computing & Technology. 69. pp. 8-12

Allen, J. (2005). "What is post-editing?" Translation Automation. 4: 1-5. Available from www.geocities.com/mtpostediting/.[Accessed June 2008].

Bowker, L. (2005). "Productivity vs Quality? A pilot study on the impact of translation memory systems". Localisation Reader 2005-2006: pp. 133-140.

Brunette, L. Gagnon, C. Hine, J. (2005). "The Grevis Project. Revise or Court Calamity". Across Languages and Cultures 6 (1). pp. 29-45

Dragsted, B. (2004). Segmentation in Translation and Translation Memory Systems. PhD Thesis. Copenhagen. Copenhagen Business School.

Gow, Francie. (2003). "Extracting useful information from TM databases." Localisation Reader 2004-2005. pp.41-44.

Guerra Martínez, L. (2003). Human Translation versus Machine Translation and Full Post-Editing of Raw Machine Translation Output. Minor Dissertation. Dublin. Dublin City University.

Krings, H. (2001). Repairing Texts: Empirical Investigations of Machine Translation Post-editing Processes. G. S. Koby, ed. Ohio. Kent State University Press.

Language Weaver. 2008. Homepage for the language automation provider. www.languageweaver.com/home.asp. [Accessed June 2008].

LISA. (2008). Homepage of the Localisation Industry Standards Association. www.lisa.org/products/qamodel/. [Accessed June 2008].

O'Brien, S. (2006). "Eye-tracking and Translation Memory Matches" Perspectives: Studies in Translatology. 14 (3). pp. 185-205.

O'Brien, S. (2007). "An Empirical Investigation of Temporal and Technical Post-Editing Effort". Translation and Interpreting Studies (tis). II, I

O'Brien, S. Fiederer, R. (2009). "Quality and Machine Translation: A Realistic Objective?". Journal of Specialised Translation, 11.

Papineni, K. Roukos, S. Ward, T. Zhu, W.J. (2002). "BLEU: A method for automatic evaluation of machine translation". In Proceedings of Association for Computational Linguistic. Philadelphia: 311-318. Also available from http://acl.ldc.upenn.edu/P/P02/P02-1040.pdf. [Accessed June 2008].

Ribas, C. (2007). Translation Memories as vehicles for error propagation. A pilot study. Minor Dissertation. Tarragona. Universitat Rovira i Virgili.

SDL. (2008). Homepage of SDL Trados 2007. www.sdl.com/en/products/products-index/sdl-trados/default.asp. [Accessed June 2008].