CHAPTER YYY

THE ROLE OF PROFESSIONAL EXPERIENCE IN POST-EDITING FROM A QUALITY AND PRODUCTIVITY PERSPECTIVE

ANA GUERBEROF ARENAS

Introduction

In this chapter, we present results on the impact of professional experience on the task of post-editing. These results are part of a larger research project where 24 translators and three reviewers were tested to obtain productivity, words per minute, and quality data, errors in final target texts, in the post-editing of machine translation (MT) and fuzzy match segments (in the 85 to 94 range). We will discuss here the results on the participants' experience according to their responses in a post-assignment questionnaire and explain how they were grouped into different clusters in order to correlate firstly the experience with speed according to the words per minute in the different match categories: Fuzzy matches, MT matches (MT output) and No match and secondly, to correlate them with the quality provided by measuring the errors marked by the three reviewers in each match category. Finally, conclusions will be drawn in relation to the experience and the resulting speed and number of errors.

Related work

There are several studies on the topic of post-editing in recent years exploring different aspects of this activity such as technical and cognitive effort: O'Brien (2006a, 2006b), Beinborn (2010) and Carl et al (2011); productivity measurement and quality: Fiederer and O'Brien (2009), Flournoy and Duran (2009), García (2010, 2011), Plitt and Masselot (2010) and De Sutter and Depraetere (2012); post-editing effort and automatic metric scores: Offersgaard et al. (2008), Tatsumi (2010) and Koponen (2012), Tatsumi and Roturier (2010), O'Brien (2011), De Sutter (2012); confidence scores: Specia (2009a. 2009b, 2011) and He et al. (2010a, 2010b), to name just a few. However, there are fewer studies exploring experience in particular and its correlation with speed and numbers of errors. We would like to mention two studies in particular. De Almeida and O'Brien (2010) explore the possible correlation between post-editing performance and years of translation experience. This pilot experiment is carried out with a group of six professional translators (three French and three Spanish) in a live localisation project using Idiom Workbench as the translation tool and Language Weaver as the MT engine. Four translators had experience in post-editing while two others did not. To analyse this performance a LISA QA Model is used in combination with the GALE post-editing guidelines. The results show that the translators with the most experience are the fastest post-editors but they also make the higher number of preferential changes. Depraetere (2010) analyses text post-edited by ten translation trainees in order to establish post-editing guidelines for translators' post-editing training. The analysis shows that students follow the instructions given and they do not rephrase the text if the meaning is clear, the students "did not feel the urge to rewrite it" (ibid: 4), they are not, however, sufficiently critical of the content thus leaving errors that should be corrected according to the instructions. Depraetere points out that this indicates a "striking difference in the mindset between translation trainees and professionals" (ibid: 6). Despite the fact that this study is focused on students, we find that it might be applicable to junior translators who have been exposed to machine translation either during their training or from the beginning of their professional experience as opposed to more senior translators that might have experienced MT at a later stage in their professional life.

Finally, we would like to mention the pilot project that served as preparation for this larger research project (Guerberof 2008) with eight subjects. In this project, we found that translators' experience had an impact on the processing speed: translators with experience performed faster on average. When we looked at the number of years of experience in localisation, domain, tools and post-editing MT output, we observed an increasing curve up to the 5-10 year range and then a drop in the speed. The number of errors was higher in experienced translators by a very small margin, and there were more errors in MT segments. This pointed to the fact that experienced translators might grow accustomed to errors in MT output. On the other hand, translators with less experience had more errors in the segments they translated from scratch than in the MT segments, which seemed to indicate that MT had a levelling effect on their quality. We felt, however, that the sample of eight participants was a highly limiting factor. It was necessary to explore further the relationship between productivity, quality and experience with a greater number of participants.

Hypothesis

Localization has a strong technical component because of the nature of the content translated as well as the tools required to translate. On many occasions this experience is associated with speed, that is, the more experience in localisation, tools used and domain, the less time will be needed to complete a project. Therefore, our hypothesis proposes that the greater the experience of the translator, the greater the productivity in post-editing MT match and Fuzzy match segments. We also formulate a sub-hypothesis that claims that this technical experience will not have an impact on the quality (measured in number of errors) as was observed in the pilot project (Guerberof 2008).

Material and method

A trained Moses (Koehn et al. 2007) statistical-base 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 resulting output obtained a BLEU score (Papineni et al. 2002) of 0.6 and a human evaluation score of 4.5 out of 5 points. The project involved the use of a web-based post-editing tool designed by CrossLang to post-edit and translate a text from English into Spanish. The file set used in the project was a new set of strings for the help system and user interface from the same customer and therefore different than the parallel data used to train the engine. It 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 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). The final output was then evaluated by three professional reviewers, who registered the errors using the LISA QA model. The focus was on the number and classification on errors, and not on a Fail or Pass result for each individual translator.

Results

As part of the global project, we analysed the 24 translators' productivity and we observed no significant differences in speed or quality for processing either the MT segments or the TM segments. Moreover, there were wide ranges in the processing speed of MT outputs so we established the possibility that some of these MT segments might have been perfect matches that required no change while others required substantial work. When looking at the impact MT had on the final quality of the post-edited text, we concluded 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 with an approved glossary, than in the MT and Fuzzy match categories. The qualitative analysis showed us that the high quality of the MT output was possibly one of the reasons for the translators showing fewer errors in the MT category than in the No match. It also showed that there were certain factors that might have influenced the translators' quality negatively: the fact that they could not go back to translated or post-edited segments, that they did not have a context for the segments, that the glossary was not integrated into the tool, that the source text contained ambiguous structures, and that the instructions might have been too vague for certain translators. These factors highlight several issues to consider when measuring quality, and when organising projects.

Finally, we analysed the data considering the translators' experience which is the focus of this chapter and we will be presenting these results in following sections.

Results on translators' experience

We are aware that the experience embraces several aspects of a translator's profile. For the purpose of this study, experience is defined as a combination of years of experience in localisation, subject matter, tools knowledge, post-editing, type of tasks performed, estimation of daily throughputs and average typing speed. The data were obtained from the questionnaire that was provided to the translators through SurveyMonkey upon completion of the assignment. The translators responded to the following questions:

- How long have you been working in the localisation industry?
- How long have you been using translation memory tools (such as SDL Trados, Star Transit, Déjà Vu)?
- How long have you been translating business intelligence software (such as SAP, Oracle, Microsoft)?
- How long have you been post-editing raw machine translated (MT) output?
- Please estimate the percentage, on average, that post-editing MT output represents in your work (considering the last three years)
- What tasks does your work involve? (You can choose more than one option).
- Please estimate your average daily throughput when you translate from scratch without any translation aid:
- What is your average typing speed? (Please, provide an estimate in words per minute).

We present a brief overview of their responses in order to understand better the experience of the participants before they are grouped into different clusters.

Answer Options	**Response %**
No experience.	**0.0%**
Less than 2 years.	**0.0%**
2 years or more, less than 4 years.	**12.5%**
4 years or more, less than 6 years.	**12.5%**
6 years or more, less than 8 years.	**25.0%**
8 years or more.	**50.0%**

Table YYY.1: Experience in the localisation and TM tools

The responses indicate that they are professional translators with experience. All translators have more than two years' experience in the localisation industry and half of them have more than eight years.

Answer Options	**Response %**
Never.	**8.3%**

Less than 2 years.	8.3%
2 years or more, less than 4 years.	4.2%
4 years or more, less than 6 years.	29.2%
6 years or more, less than 8 years.	16.7%
8 years or more.	33.3%

Table YYY.2: Experience in domain

The experience is more heterogeneous in this group in relation to the domain, business intelligence translation, but still only four translators have less than two years' experience or none.

Answer Options	Response %
Never.	25.0%
Less than 2 years.	29.2%
2 years or more, less than 4 years.	25.0%
4 years or more, less than 6 years.	8.3%
6 years or more, less than 8 years.	4.2%
8 years or more.	8.3%

Table YYY.3: Experience in post-editing

The responses show that post-editing is a relatively new task for the translators in comparison with their experience in the other areas, 79.2 percent has no experience or less than four years' experience on the task.

Answer Options	Response %
0%	25.0%
1% to 25%	66.7%
26% to 49%	4.2%
50% to 74%	4.2%
75% to 90%	0.0%
91% to 100%	0.0%

Table YYY.4: Estimated post-editing work in the last three years

We wanted to qualify the previous questions as some translators might have certain experience in post-editing but they might not perform it on a regular basis and we can see on Table YYY.4, rows 1 and 2, that post-editing still does not represent a high percentage of work for them.

Tasks		No		Yes
Post-editing		**37.50**		**62.50**
Translating		**4.17**		**95.83**
Revising		**12.50**		**87.50**
Writing		**83.33**		**16.67**
Terminology work		**62.50**		**37.50**
Other		**79.17**		**20.83**

Table YYY.5: Tasks performed

The 24 translators are more focused on translating and revising activities.

Answer Options	Response %
Less than 2000 words per day.	**8.3%**
Between 2100 and 3000 w/ per day.	**70.8%**
Between 3100 and 5000 w/ per day.	**20.8%**
More than 5100 words per day.	**0.0%**
I don’t know	**0.0%**

Table YYY.6: Estimated daily throughput

The majority selected the option between 2,100 and 3,000 words per day which is considered a standard metric in the industry and thus not surprising.

Answer Options	Response %
0-20 words per minute	**8.3%**
21-40 words per minute	**16.7%**
41-60 words per minute	**41.7%**
61-80 words per minute	**20.8%**
More than 81 words per minute	**12.5%**

Table YYY.7: Estimated typing speed

All responses suggest that this is a group of 24 professional translators with different areas of expertise, although there are three translators with considerable less experience than the remaining twenty-one. Most have experience using tools and some experience in post-editing MT output, although the task represents a low percentage of their work and has not been performed for a very long period of time. Finally, their working speed seems to be in accordance with the industry standard. Now, we should look into how these translators were grouped into clusters to test the hypothesis.

Grouping translators according to their experience

In order to distribute translators into different groups with similar experience, a multiple correspondences analysis was setup (Greenacre 2008). This enables us to represent all the data (responses from the questionnaire by all translators) as rows and columns in a table including active variables (the questions above) and showing illustrative variables (age and sex). These were then graphically represented as dots in a two dimensional map (biplot). Four groups (clusters) were found, with distinctive characteristics. To explain the complete statistical analysis is beyond the scope of this study, but we should mention that the factors are not pre-defined, as we plot the data to see how the different variables are related in order to understand their relation and hence define the clusters.

We obtained four clusters that are characterised as follows. Cluster 1 has experience in all the areas queried, but they have been doing these tasks for a shorter period of time than those in Cluster 2. The translators in this cluster have between six and eight years' experience in localisation and TM tools, between four and six years' experience in translating business intelligence and 50 percent of them have a speed ranging from 21 to 60 words per minute. Cluster 2 is the one with the most experience. The translators in this cluster have more than eight years' experience in the localisation industry, more than eight years' experience using TMs, more than eight years' experience in translating business intelligence and all translators in this cluster work in post-editing. Cluster 3 has experience in translation, but none or less experience in post-editing MT output. Finally, Cluster 4 is characterised by being young and having less professional experience. Both translators in this cluster have less than two years' experience translating business intelligence and they are less than 25 years old.

Experience vs. processing speed: Fuzzy match

The speed (words per minute) for the Fuzzy match segments processed by the translators is calculated taking the words per minute in Fuzzy match segments according to the translators' different clusters:

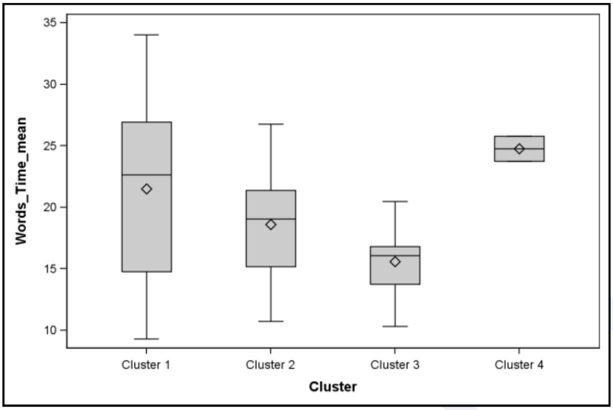

Figure YYY.1: Processing speed in words per minute vs. Fuzzy match

Cluster 3, with no or little experience in post-editing, shows lower processing speed in Fuzzy match than the other clusters. Cluster 1, the second in overall experience, has a higher mean and median values than Clusters 2 and 3. Cluster 2, the most experienced, behaves similarly to Cluster 1 but slower than Cluster 4, which has a very homogeneous speed (only two translators) and the highest mean and median values. Let us look at the descriptive data in Table YYY.8.

Cluster	Min	Median	Mean	Max	SD
1	9.29	22.65	21.49	34.03	8.41
2	10.73	19.05	18.59	26.74	4.95
3	10.33	16.07	15.58	20.48	3.37
4	23.75	24.76	24.76	25.78	1.43

Table YYY.8: Processing speed vs. Fuzzy match

Cluster 1 has the second highest mean and median values with the highest deviation. Cluster 2 has slightly lower figures. Cluster 3 has the lowest values. Cluster 4 has the highest mean and median values and is the most homogenous group.

Therefore, if Fuzzy matches are examined in the clusters with more experience (1 and 2) the productivities are high. However, productivities are also high in Cluster 4, the one with the least experience. The interesting data point in this case is that Cluster 3, with no or little experience in post-editing, although with experience on the other areas, has a lower processing speed than the other three clusters. This might indicate that this particular cluster was slower when processing the data because their typing speed was slower (the two slowest typists are in this cluster) or because they invested more time in producing a better translation (we will see this in the following section when we look at the errors per cluster). But how did the clusters then behave with MT matches? Was this Cluster 3, with no experience in post-editing, also the slowest in this category?

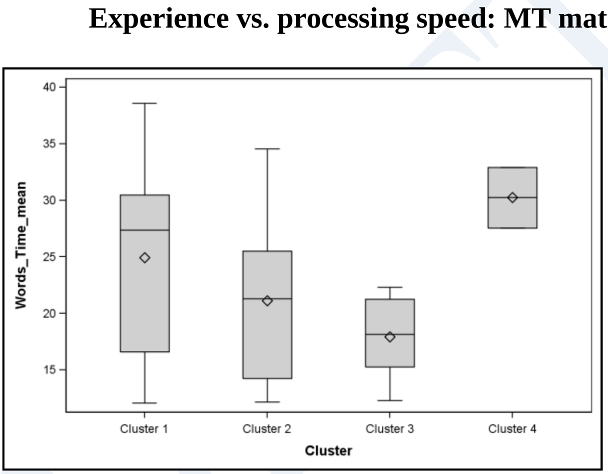

Figure YYY.2: Processing speed in words per minute vs. MT match

Cluster 4, with the least experience, seems to have taken full advantage of MT matches, with very high median and mean. Cluster 1 and Cluster 2, with the most experience, show similar values, although Cluster 1 seems to be slightly faster. There are translators in Clusters 1 and 2 that seem to have quite different speeds. Cluster 3, with no post-editing experience, has more homogenous values and again the lowest mean and median values. This might be understandable if they declare having no experience in post-editing MT. Let us look at the descriptive data (Table YYY.9) to gain better understanding of the figures above.

Cluster	Min	Median	Mean	Max	SD
1	**12.07**	**27.38**	**24.94**	**38.58**	**9.09**

2	**12.17**	**21.29**	**21.10**	**34.57**	**7.57**
3	**12.31**	**18.14**	**17.90**	**22.33**	**3.82**
4	**27.55**	**30.23**	**30.23**	**32.91**	**3.79**

Table YYY.9: Processing speed vs. MT match

Cluster 4 clearly has high processing speeds when dealing with MT matches. Cluster 3 has the lowest values if the mean and median values are considered, there is a maximum speed of 22.33 words per minute, the deviation here being lower than in Clusters 1 and 2. Clusters 1 and 2 have similar minim and maximum values, although Cluster 1 shows faster mean and median values.

If Cluster 4 shows the highest mean and median values, it seems to show quite the opposite of what we were trying to test. These translators are young and have very little experience but they seem to benefit considerably from MT. Nevertheless, we also see that specific experience could be a factor. Cluster 3, the slowest, had no or little post-editing experience. This seems to indicate that younger translators might find it easier to deal with MT post-editing because they might have had more contact with MT or TM outputs since they started working professionally (we saw, when defining the clusters, that these two translators had the same experience in localisation as in post-editing, which shows that they have almost a parallel experience in both areas, while more senior translators do not). At any rate, Clusters 1 and 2, with more experience, still have the highest values at 38.58 and 34.57 in words per minute respectively. Overall experience can have different influences. On the one hand, translators with more experience can perform well, and on the other, translators with less experience can also make good use of MT segments (possibly if exposed to or trained in machine translation post-editing).

It will be interesting to see how these four clusters perform when translating on their own, to find out if the different productivities were also related to their own (intrinsic) speed in No match words.

Experience vs. processing speed: No match

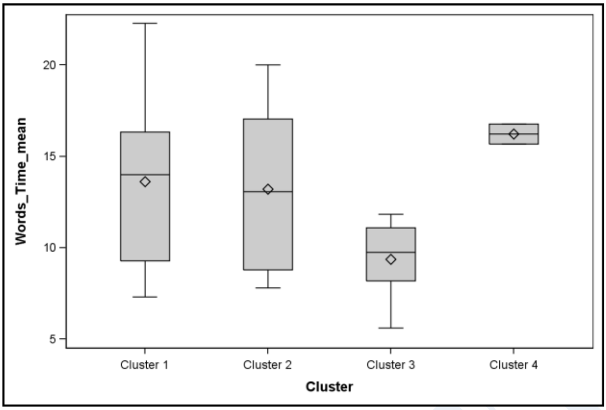

Figure YYY.3: Processing speed in words per minute vs. No match

Cluster 4 has the highest mean and median values for the No match category. These two translators seem to work at a reasonable speed also when working without a translation aid. Cluster 1 is the second fastest in mean and median values and also seems to have the maximum value in words per minute. Cluster 2 has similar values with a wider range in the quartiles than Cluster 1. Cluster 3 is the group with the lowest mean and median values, and also includes the translator with the lowest value in all the clusters. Table YYY.10 shows the descriptive data.

Cluster	Min	Median	Mean	Max	SD
1	7.32	14.00	13.61	22.29	5.00
2	7.80	13.08	13.20	20.00	4.70
3	5.60	9.75	9.37	11.85	2.24
4	15.69	16.24	16.24	16.78	0.77

Table YYY.10: Processing speed vs. No match

Cluster 4 has the highest processing speeds if we look at the median and mean values, and also less deviation (only two translators). However, Cluster 1 has the maximum value followed by Cluster 2. The translators in Cluster 3 present lower values overall but less deviation that shows more homogeneity in the translators’ speeds.

It seems understandable that Cluster 3 also had low processing speeds when working with MT and Fuzzy matches, since their baseline (No match translation) is within a low speed range. It is, therefore, not clear if their low productivity in the three match categories (Fuzzy, MT and No match) was due to their speed as translators, to lack of experience in post-editing MT output (the lack of familiarity with these types of errors might decrease their speed) or simply because they had spent more time in correcting errors. It is also interesting to note that all the translators that declare having an average typing speed of 0-20 words per minute are in this cluster.

By looking at the descriptive data it is difficult to know if experience made a statistically significant difference in processing speed. A linear regression model with repeated measures was applied to the data, taking logarithm of *Words per minute* as the response variable, and *Match category* and *Cluster* as explanatory variables. There are statistically significant differences (F=169.91 and p<0.0001) between the three translation categories: Fuzzy match, MT match and No match. This is exactly what we saw when we analysed productivity. However, there are no statistically significant differences between Clusters, and in the interaction between Clusters and Match category. From this model, mean value estimations were calculated taking the variable logarithm of *Words per minute according to the Match and Cluster*. We present the estimated mean value with their corresponding confidence intervals of 95 percent. The estimations are expressed in words per minute for a better understanding.

Cluster	**Mean**	**Lower**	**Upper**
Cluster 1	**18.09**	**14.27**	**22.91**
Cluster 2	**16.46**	**12.99**	**20.86**
Cluster 3	**13.46**	**10.24**	**17.69**
Cluster 4	**22.95**	**14.30**	**36.84**

Table YYY.11: Estimated mean in words per minute per Cluster

Although the estimated mean for Cluster 4 is the highest, followed by Clusters 1, 2 and Cluster 3, there are no statistically significant differences between the four clusters. The gap between Cluster 3 and Cluster 4 is approximately 9 words. The lower and upper intervals overlap with each other, showing that the translators in each cluster presented a variety of speeds not necessarily related to experience. This is contrary to the findings from De Almeida and O’Brien (2010) and our pilot project (Guerberof 2008) where faster translators were also the ones with more experience. However, the number of participants was smaller, and this made it difficult to see the effect experience had on speed. Table YYY. 12 shows the estimated mean again, but now showing the Match category and the Productivity gain with respect to No match.

Match	Cluster	Estimated mean	L	U
Fuzzy	**1**	**19.85**	**15.56**	**25.32**
Fuzzy	**2**	**17.98**	**14.09**	**22.93**
Fuzzy	**3**	**15.26**	**11.52**	**20.21**
Fuzzy	**4**	**24.74**	**15.21**	**40.26**
MT	**1**	**23.31**	**18.28**	**29.74**
MT	**2**	**19.94**	**15.63**	**25.44**
MT	**3**	**17.54**	**13.24**	**23.24**
MT	**4**	**30.11**	**18.51**	**49.00**
No match	**1**	**12.79**	**10.02**	**16.31**
No match	**2**	**12.45**	**9.76**	**15.88**
No match	**3**	**9.11**	**6.88**	**12.07**
No match	**4**	**16.23**	**9.97**	**26.40**

Table YYY.12: Estimated mean according to Match and Cluster

Speed is always lower for Cluster 3, higher for Cluster 4, and similar for Clusters 1 and 2 in the three match categories. No match is significantly different for all clusters, while Fuzzy match and MT match show similar values, except with Cluster 4, where the MT match is slightly higher. To double-test the validity of the findings, non-parametric comparisons were set-up (Kruskal-Wallis analysis of variance) and we found no statistically significant differences between the Clusters according to the Match category if speed was considered.

Consequently, the first part of our hypothesis that says that the greater the experience of the translator, the greater the productivity in post-editing MT match and Fuzzy match segments is not supported in our experiment. Although Clusters 1 and 2, with more experience, show high values, Cluster 4, with less experience, also shows the highest mean and median results. Cluster 3, on the other hand, with no post-editing experience, shows lower speed values, but this was also the case in the No match category. Hence the reason could lie more in their own average typing speed or general processing speed than in the fact that they have no experience in post-editing MT matches.

In the same way that productivity needs to be linked to quality, experience needs to be related to productivity and to quality. Would Cluster 4 present more errors than Cluster 3, for example?

Experience vs. number of errors: Fuzzy matches

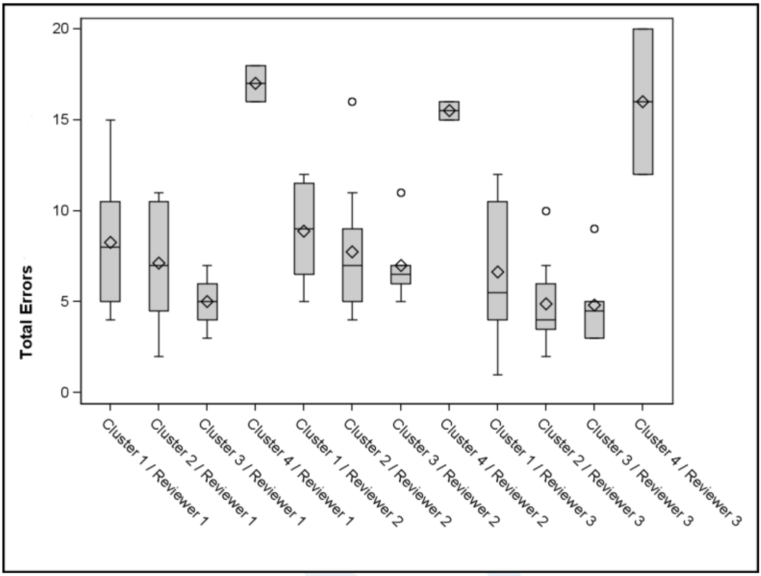

Figure YYY.5: Total errors for Fuzzy match in clusters

Interestingly, Cluster 4 has the highest number of errors according to all three reviewers, indicating that this Cluster was the fastest if the mean value is considered, but it was not as rigorous or thorough when editing the Fuzzy match category. On the other hand, Cluster 3 has the lowest number of errors, indicating that this Cluster was the slowest but also thorough when processing the Fuzzy match segments. The differences between Clusters 1 and 2 are not pronounced.

Cluster & Rev		**N**	**Mean**	**Median**	**SD**	**Min**	**Max**
1	**Rev 1**	**8**	**8.25**	**8.00**	**3.77**	**4**	**15**
	Rev 2	**8**	**8.88**	**9.00**	**2.90**	**5**	**12**
	Rev 3	**8**	**6.63**	**5.50**	**3.96**	**1**	**12**
2	**Rev 1**	**8**	**7.13**	**7.00**	**3.48**	**2**	**11**
	Rev 2	**8**	**7.75**	**7.00**	**3.99**	**4**	**16**

Cluster & Rev		N	Mean	Median	SD	Min	Max
	Rev 3	8	4.88	4.00	2.53	2	10
3	Rev 1	6	5.00	5.00	1.41	3	7
	Rev 2	6	7.00	6.50	2.10	5	11
	Rev 3	6	4.83	4.50	2.23	3	9
4	Rev 1	2	17.00	17.00	1.41	16	18
	Rev 2	2	15.50	15.50	0.71	15	16
	Rev 3	2	16.00	16.00	5.66	12	20

Table YYY.13: Total errors for Fuzzy match in clusters

Cluster 4 has the highest mean values for all three reviewers, the highest median values, and the highest minimum and maximum values. The only similar maximum value is in Cluster 2. Cluster 3 has the lowest mean and median values from the three reviewers. However, the minimum and maximum values are very similar in these three clusters (1, 2 and 3), indicating that some translators had low or high values irrespective of the cluster they were in. When the type of errors is consulted, Cluster 4 made more mistakes in Terminology. This clearly indicates that translators in Cluster 4 gained speed because they tended not to check the glossary. They accepted the terminology as it was presented to them in the Fuzzy matches. We observe that Cluster 3 was slowest because they might have devoted more time to check the terminology against the glossary provided.

For Fuzzy matches, the results are rather clear. Cluster 4, with less experience and higher speed, left or made more errors in the segments according to the three reviewers. Cluster 3 made slightly less, although results for Clusters 1, 2 and 3 are quite similar. These results are interesting since they seem to signal a lack of attention to certain important aspects of the translation process in the more novice translators. We suspect that this would be the case for the whole assignment, but let us have a look at the results for the MT matches in Figure YYY.6.

Experience vs. number of errors: MT matches

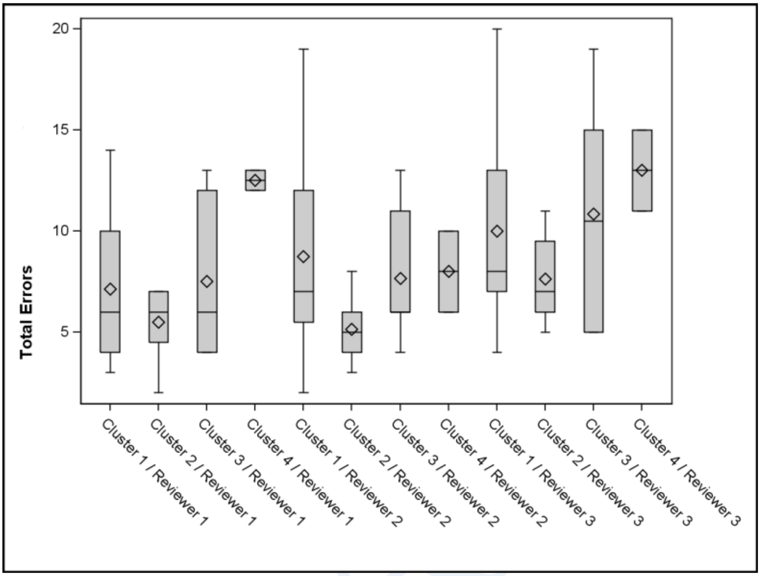

Figure YYY.6: Total errors for MT match in clusters

These results are particularly interesting. In this case, the differences between the clusters are not as pronounced as with the Fuzzy matches. We think this is possible because some of the MT matches were perfect matches, with no changes required, and although translators can still introduce mistakes, it would be logical that if the translators in Cluster 4 had problems in terminology (failing to check the glossary consistently, and a certain lack of understanding of instructions), the perfect matches could help them lower the number of errors. Table YYY.14 shows the descriptive data for MT match.

Cluster & Reviewer		**N**	**Mean**	**Median**	**SD**	**Min**	**Max**
1	**Rev 1**	**8**	**7.13**	**6.00**	**4.19**	**3**	**14**
	Rev 2	**8**	**8.75**	**7.00**	**5.60**	**2**	**19**
	Rev 3	**8**	**10.00**	**8.00**	**5.21**	**4**	**20**
2	**Rev 1**	**8**	**5.50**	**6.00**	**1.77**	**2**	**7**
	Rev 2	**8**	**5.13**	**5.00**	**1.64**	**3**	**8**
	Rev 3	**8**	**7.63**	**7.00**	**2.13**	**5**	**11**
3	**Rev 1**	**6**	**7.50**	**6.00**	**4.04**	**4**	**13**

Cluster & Reviewer		N	Mean	Median	SD	Min	Max
	Rev 2	**6**	**7.67**	**6.00**	**3.50**	**4**	**13**
	Rev 3	**6**	**10.83**	**10.50**	**5.60**	**5**	**19**
4	**Rev 1**	**2**	**12.50**	**12.50**	**0.71**	**12**	**13**
	Rev 2	**2**	**8.00**	**8.00**	**2.83**	**6**	**10**
	Rev 3	**2**	**13.00**	**13.00**	**2.83**	**11**	**15**

Table YYY.14: Total errors for MT match in clusters

Cluster 2 has the lowest mean values and Cluster 4 the highest if we consider all three reviewers. However, not all the values are as different as what we saw in the Fuzzy match category. Cluster 4 has the highest minimum values, but the maximum values are to be found in Cluster 1. If we look at the type of errors each Cluster made the results are different from those found in Fuzzy matches. There are Terminology errors but here the majority of errors are on Language overall, according to all three reviewers. The reviewers seem to be of the opinion that not enough changes were made in the segments for them to be linguistically acceptable. Still the least experienced translators did not check the glossary with MT matches because they have almost an equal number of Terminology errors. Cluster 2, the most experienced, performed better with MT matches with fewer errors and fewer Language errors than the other clusters. Hence, this might indicate that experience is a factor when dealing with MT matches in terms of quality, but also that the differences in errors between the clusters were not as pronounced as in Fuzzy matches. Cluster 4 performed faster with MT matches and the number of errors was lower than with Fuzzy matches, and this might indicate that with translators who have less experience, high quality output MT might be a better option than translation memories below the 94 percent threshold.

If translators behave differently with Fuzzy than with MT matches, how did they do without any translation proposal? Figure YYY.7 shows the results for the No match category.

Experience vs. number of errors: No matches

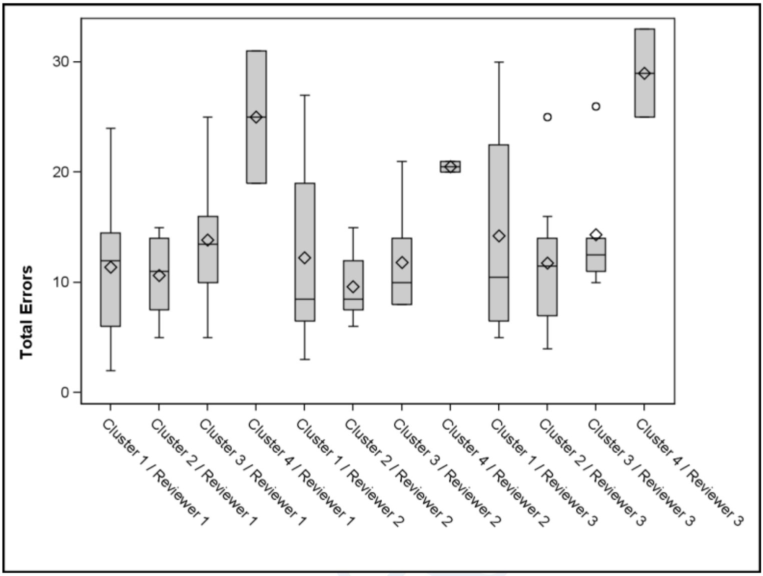

Figure YYY.7: Total errors for No match in clusters

The results here are more similar to the Fuzzy match than to the MT match results. Cluster 4 clearly has the highest number of errors, and the other three clusters are very close in results. Once again, Cluster 2 seems to have the most homogenous data, thus indicating that this cluster did not have translators with extreme values as in Clusters 1 and 3. Table YYY.15 shows the descriptive values for the No match category.

Cluster & Reviewer		**N**	**Mean**	**Median**	**SD**	**Min**	**Max**
1	**Rev 1**	**8**	**11.38**	**12.00**	**6.86**	**2**	**24**
	Rev 2	**8**	**12.25**	**8.50**	**8.38**	**3**	**27**
	Rev 3	**8**	**14.25**	**10.50**	**9.68**	**5**	**30**
2	**Rev 1**	**8**	**10.63**	**11.00**	**3.93**	**5**	**15**
	Rev 2	**8**	**9.63**	**8.50**	**3.11**	**6**	**15**

Cluster &	Reviewer	N	Mean	Median	SD	Min	Max
	Rev 3	8	11.75	11.50	6.56	4	25
3	Rev 1	6	13.83	13.50	6.68	5	25
	Rev 2	6	11.83	10.00	5.04	8	21
	Rev 3	6	14.33	12.50	5.89	10	26
4	Rev 1	2	25.00	25.00	8.49	19	31
	Rev 2	2	20.50	20.50	0.71	20	21
	Rev 3	2	29.00	29.00	5.66	25	33

Table YYY.15: Total errors for No match in clusters

Cluster 4 clearly has the highest mean and median values according to all reviewers. They also have a very high minimum value. Cluster 3 has higher aggregated values, but all three clusters have similar median and mean values, showing that many translators have similar numbers of errors. If we look at the type of errors each Cluster made, the results are slightly different from those found for Fuzzy and MT matches. The majority of errors are in Language, followed by Terminology and Style. The reviewers seem to be of the opinion that the segments were not linguistically acceptable, as with MT matches. However, when we look at Cluster 4, the majority of errors are in Terminology. Once again, the glossary and the instructions were not followed correctly. The number of errors in Clusters 1, 2 and 3 are similar. This seems to point to the fact that translators with experience work better with the instructions given and are more thorough. This was also true for Fuzzy matches and to a lesser extent for MT matches.

Are these differences significant? We saw differences in speed but these were not statistically significant between the Clusters, so what will be the case for the number of errors? A Poisson regression model is applied with repeated measures taking the variable *Total errors* as the response variable and the offset as text length. Statistically significant differences are observed for the variable *Total errors* between the different Match categories: Fuzzy, MT and No match (F=53.50 and p<0.0001), as well as for the different clusters (F=7.61 and p<0.0001). Finally, statistically significant differences are observed in the interaction between Match categories and Clusters (F=3.37 and p=0.0039).

From this model, estimations of the mean values are obtained for the variable (total errors /text length) according to Match category with the corresponding interval levels of 95 percent. We present the results of these estimations but expressed in number of errors per segment length for better understanding. We consider the length of the original text (Fuzzy match, 618 words, MT match, 757 words and No match 749 words).

Match	Cluster	Mean	SD	L	U
Fuzzy	**1**	**7.41**	**0.74**	**6.08**	**9.03**
Fuzzy	**2**	**6.41**	**0.67**	**5.21**	**7.89**
Fuzzy	**3**	**5.42**	**0.69**	**4.21**	**6.96**
Fuzzy	**4**	**16.04**	**2.72**	**11.47**	**22.44**
MT	**1**	**8.07**	**0.79**	**6.65**	**9.79**
MT	**2**	**5.93**	**0.64**	**4.79**	**7.33**
MT	**3**	**8.37**	**0.94**	**6.70**	**10.45**
MT	**4**	**11.08**	**2.02**	**7.72**	**15.90**
New	**1**	**11.81**	**1.06**	**9.89**	**14.10**
New	**2**	**10.39**	**0.96**	**8.65**	**12.48**
New	**3**	**12.87**	**1.31**	**10.52**	**15.75**
New	**4**	**24.65**	**3.91**	**18.01**	**33.73**

Table YYY.16: Estimated mean of errors per match and cluster

When we observe the interaction between Clusters and Match categories in Table YYY.16, the results are interesting once again. Cluster 4 shows statistically significant differences in the Fuzzy match and No match categories. But in the MT match category, although the number of errors is higher, the confidence intervals overlap (row 8), showing that this difference is not statistically significant in this particular match category. So MT, in this instance, acted as a “leveller” in terms of errors for Cluster 4. The results are in line with the findings from De Almeida and O’Brien (2010) where more experienced translators were more accurate and also with Guerberof (2008) where MT had a levelling effect with novice translators.

The second part of our hypothesis claims that experience will not have an impact on the quality (measured in number of errors). Now, after going through the results, we find that this hypothesis is not supported by our data. In fact, the results show the opposite, that experience does play a part in the number of errors found. It is true that for Clusters 1, 2 and 3 there are no statistically significant differences, but there are for Cluster 4 that represented the novice group. The translators made more mistakes, mainly because they did not follow instructions and hence avoided the glossary, resulting in a higher speed but poorer quality. Interestingly, the number of errors was not as high in MT match segments, and this could be because some segments in MT required little change or because the terminology was already consistent with the glossary. Cluster 2, the most experienced, has fewer errors although these were not significantly lower. Cluster 3, with no experience in post-editing, performed worse in this category, showing again that training and experience in this task might help not only with respect to speed but also in quality.

Conclusions on the translators' experience

All the translators are professional translators who have varying experience in localisation and using tools and some experience in post-editing MT output, although the task represents a low percentage of their work and has not been performed for a very long period of time. Their working speed seems to be in accordance with industry standards and is quite homogeneous. A multivariate analysis was setup to distribute the translators into four different clusters to test our hypothesis. The results indicate that the incidence of experience on the processing speed is not significantly different. Translators with more experience performed similarly to other very novice translators. Translators with less or no experience in post-editing were the slowest cluster but again the differences were not significant. This seems to be different from our previous findings (Guerberof 2008) and from the findings by De Almeida and O'Brien (2010), although more in line with the findings in Tatsumi (2010). However, the numbers of participants in those studies are lower, to the extent that one post-editor has a great impact in the whole group, whereas in this project there were 24 translators. Further research is needed to draw definitive conclusions.

Our findings on errors are in line with those in De Almeida and O'Brien (2010). Translators with more experience made fewer mistakes than those with less experience. As Offersgaard et al. (2008) suggests a "good post-editor is an experienced proof-reader" (ibid: 156). The number of errors was significantly different between Cluster 4 (the novice group) and the other clusters with regards to Fuzzy and No match. The difference was higher but not significant for MT match. Also the type of errors made by the novice translators were mostly Terminology errors, as opposed to Language or Style as in the other clusters, indicating that these translators with less experience were less thorough with terminology and with instructions than were the more experienced ones. But this is not to say that they did not have more errors in the other categories as well. The MT output, however, seems to have had a levelling effect as far as errors is concerned. This might lead us to suggest that using high-quality MT output as opposed to Fuzzy matches below the 95 percent threshold might be advisable for translators with less experience, as there are more probabilities of having perfect matches in the proposed texts and hence of making fewer mistakes. Are novice translators more tolerant to errors in quality than senior translators? Our reviewers were senior translators and they might have a different idea of quality than the novice translators. Is the current review method adequate to establish a quality suitable for the market? Lagoudaki (2008) and Flournoy and Duran (2009) also suggest that inexperienced translators seem to be more tolerant of MT errors and structures than experienced ones. Similarly, Depraetere (2010) pointed out that translation trainees are more tolerant of MT errors. It might be that "new" generations of translators might have a different outlook on translation quality to that of senior translators. Finally, it was also observed that the cluster with the least or no experience in post-editing performs better with Fuzzy matches in terms of errors than with MT matches, and this seems to indicate that experience and training on post-editing might have a pay-off in terms of quality, although this might not be the only factor.

Bibliography

Beinborn, L. 2010. “Post-editing of Statistical Machine Translation: A Crosslinguistic Analysis of the Temporal, Technical and Cognitive Effort.” Master of Science Thesis, Saarland University.

Carl, M., B. Dragsted, J. Elming, D. Hardt, and A. Jakobsen 2011. “The Process of Post-editing: a Pilot Study.” Paper presented in 8th international NLPSC workshop, Frederiksberg, Copenhagen, August 20-21. Accessed June 2013. http://www.mt-archive.info/NLPCS-2011-Carl-1.pdf.

De Almeida, G. and S. O’Brien 2010. “Analysing Post-Editing Performance: Correlations with Years of Translation Experience.” Paper presented at the *14th Annual Conference of the EAMT*, St. Raphael, May 27-28. Accessed June 2013. http://www.mt-archive.info/EAMT-2010-Almeida.pdf.

De Sutter, N. 2012. “MT Evaluation Based on Post-editing: a Proposal.” In *Perspectives on translation quality*, edited by Ilse Depraetere, 125-143. Berlin: Mouton de Gruyter.

De Sutter, N., and I. Depraetere 2012. “Post-edited Translation Quality, Edit Distance and Fluency Scores: Report on a Case Study.” Paper presented in *Journée d'études Traduction et qualité Méthodologies en matière d'assurance qualité*, Universtité Lille 3. Sciences humaines et sociales, Lille, February 3rd. Accessed June 2013. http://stl.recherche.univ-lille3.fr/colloques/20112012/DeSutter&Depraetere_2012_02_03.pdf.

Depraetere, I. 2010. “What Counts as Useful Advice in a University Post-editing Training Context? Report on a case study.” Paper presented at the *14th Annual EAMT Conference*, St. Raphael, May 27-28. Accessed June 2013. http://www.mt-archive.info/EAMT-2010-Depraetere-2.pdf.

Flournoy, R., and C. Duran 2009. “Machine Translation and Document Localization at Adobe: From Pilot to Production.” Paper presented at *MT Summit XII*, Ottawa, August 26-30. Accessed June 2013. http://www.mt-archive.info/MTS-2009-Flournoy.pdf.

García, I. 2010. “Is Machine Translation Ready Yet?” *Target*. Vol. (22-1): 7-21. Amsterdam and Philadelphia: Benjamins.

García, I. 2011. “Translating by Post-editing: Is it the Way Forward?” *Machine Translation*. Vol. 25(3): 217-237, Netherlands: Springer

Greenacre, M. 2008. *La práctica del análisis de correspondencias*. (Spanish translation of *Correspondence Analysis in Practice*, Second Edition). Madrid: Manuales Fundación BBVA.

Guerberof, A. 2008. “Productivity and Quality in Machine Translation and Translation Memory outputs.” Masters Dissertation, Universistat Rovira i Virgili. http://bit.ly/1a83G9p

He, Y., Y. Ma, J. Roturier, A. Way, and J. van Genabith 2010a. “Bridging SMT and TM with Translation Recommendation.” Paper presented at *48th Annual Meeting of ACL*, Uppsala, July 10-16. Accessed June 2013. http://www.mt-archive.info/ACL-2010-He.pdf.

He, Y., Y. Ma, J. Roturier, A. Way, and J. van Genabith 2010b. “Improving the Post-editing Experience using Translation Recommendation: A User Study.” Paper presented at the *9th Annual AMTA Conference*, Denver, October 31-November 4. Accessed June 2012. http://doras.dcu.ie/15803/

Koehn, P., H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst 2007. "Moses: Open Source Toolkit for Statistical Machine Translation." In *Proceedings of the Demo and Poster Sessions of the 45th Annual Meeting of the Association for Computational Linguistics (ACL ’07)*, 177–180. Prague, Czech Republic.

Lagoudaki, E. 2008. “The value of Machine Translation for the Professional Translator.” Paper presented at the *8th AMTA Conference*. Hawaii: 262-269. Accessed June 2013. http://www.amtaweb.org/papers/3.04_Lagoudaki.pdf

Koponen, M. 2012. “Comparing Human Perceptions of Post-editing Effort with Post-editing Operations.” Paper presented at the *7th Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Montreal, June 7-8. Accessed June 2013. http://www.aclweb.org/anthology-new/W/W12/W12-3123.pdf

O’Brien, S. 2006a. “Methodologies for Measuring Correlations between Post-editing Effort and Machine Translatability.” *Machine Translation:* 37-58. Netherlands: Springer.

O’Brien, S. 2006b. “Eye-tracking and Translation Memory Matches” *Perspectives: Studies in Translatology*. 14 (3): 185-205

O'Brien, S. 2011. "Towards Predicting Post-editing Productivity." *Machine Translation*. Vol. 25(3): 197-215. Netherlands: Springer.

O'Brien, S. 2012. "Towards a Dynamic Quality Evaluation Model for Translation." *Journal of Specialised Translation*. (17) Accessed June 2013. http://www.jostrans.org/issue17/art_obrien.pdf

Offersgaard, L., C. Povlsen, L. Almsten, and B. Maegaard 2008. "Domain Specific MT Use." Paper presented at the *12th EAMT Conference*. Hamburg, September 22-23. Accessed June 2013 http://www.mt-archive.info/EAMT-2008-Offersgaard.pdf

Papineni, K., S. Roukos, T. Ward, and W.J. Zhu 2002. "BLEU: A Method for Automatic Evaluation of Machine Translation." Paper presented at the *40th Annual Meeting of the Association for Computational Linguistics* (ACL), Philadelphia, July 7-12. Accessed June 2013. http://acl.ldc.upenn.edu/P/P02/P02-1040.pdf.

Plitt, M., and F. Masselot 2010. "A Productivity Test of Statistical Machine Translation Post-editing in a Typical Localisation Context." *The Prague Bulletin of Mathematical Linguistics*. Prague: 7-16. Accessed June 2013. http://ufal.mff.cuni.cz/pbml/93/art-plitt-masselot.pdf.

Snover, M., B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul 2006. "A Study of Translation Edit Rate with Targeted Human Annotation." Paper presented at *7th Annual AMTA Conference*, Cambridge, August 8-12. Accessed June 2013. http://mt-archive.info/AMTA-2006-Snover.pdf.

Specia, L. 2011. "Exploiting Objective Annotations for Measuring Translation Post-editing Effort." Paper presented at 15th Annual EAMT Conference, Leuven, May 30-31. Accessed June 2013. http://www.mt-archive.info/EAMT-2011-Specia.pdf.

Specia, L., N. Cancedda, M. Dymetman, M. Turchi, and N. Cristianini 2009a. "Estimating the Sentence-level Quality of Machine Translation Systems." Paper presented at the *13th Annual Conference of the EAMT*, Barcelona, May 14-15. Accessed June 2013. http://clg.wlv.ac.uk/papers/Specia_EAMT2009.pdf.

Specia, L., C. Saunders, M. Turchi, Z. Wang, and J. Shawe-Taylor 2009b. "Improving the Confidence of Machine Translation Quality Estimates." Paper presented at *MT Summit XII*, Ottawa, August 26-30. Accessed June 2013. http://eprints.pascal-network.org/archive/00005490/01/MTS-2009-Specia.pdf.

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/

Tatsumi, M., and J. Roturier 2010. "Source Text Characteristics and Technical and Temporal Post-editing Effort: What is their Relationship?" Paper presented at the *2nd Joint EM+/CNGL workshop "Bringing MT to the user: research on integrating MT in the translation industry",* Denver, November 4th. Accessed June 2013. http://www.mt-archive.info/JEC-2010-Tatsumi.pdf.

Turian, J., L. Shen, and I.D. Melamed 2003. "Evaluation of Machine Translation and Its Evaluation." Paper presented at *MT Summit IX*, New Orleans, September 23-27. Accessed June 2012. http://nlp.cs.nyu.edu/pubs/papers/turian-summit03eval.pdf.