Quality is in the eyes of the reviewer:
A report on post-editing quality evaluation

Ana Guerberof Arenas

Abstract

As part of a larger research project exploring correlations between productivity, quality and experience in the post-editing of machine-translated and translation-memory outputs in a team of 24 professional translators, three reviewers were asked to review the translations/post-editions completed by these translators and to fill in the corresponding quality evaluation forms. The data obtained from the three reviewers' evaluation was analyzed in order to determine if there was agreement in terms of time as well as in number and type of errors marked to complete the task. The results show that there were statistically significant differences between reviewers, although there were also correlations on pairs of reviewers depending on the provenance of the text analyzed. Reviewers tended to agree on the general number of errors found in the No match category but their agreement in Fuzzy and MT match was either weak or there was no agreement, perhaps indicating that the origin of the text might have influenced their evaluation. The reviewers also tended to agree on best and worst performers, but there was great disparity in the translators' classifications if they were ranked according to the number of errors.

Keywords

Revision; post-editing; MT; translation quality; reviewers; errors; TM

Introduction

The review phase is an important step in the localization process as most quality control methods and standards testify. This review phase is generally defined as a revision carried out by a person other than the translator (European quality standard, EN15038) with the purpose of assessing the quality or suitability of the translation according to a specific project briefing. Reviews are also carried out to recruit new free-lance translators, assess novice translators, and give feedback to translators in a project, and most recently to assess the quality of machine translation output and of post-edited texts. Reviewers are usually senior translators themselves specialized in a specific area of knowledge, a product, or even a particular customer. Since the translation buyers cannot possibly be proficient in all languages, reviewers often decide which translators should be part of a team, if the quality of a translation is suitable for a release, if an agency's performance is adequate, or if machine-translated output has the sufficient quality to be used as part of a given localization project.

Defining how to assess this quality, however, is not an easy task. If we browse through topics in localization conferences, quality is one of those always present in the program. Quality is a difficult concept to define because it needs to adapt to a fast pace, quick changing and financially challenged "industry". This accounts for the volume of existing literature on the topic. Even if defining quality, and how to measure it, is a hard enterprise in an industry marked by an increasing level of complexity, most language service providers (LSPs) continue to measure quality according to the number of errors found in the final target text (O'Brien 2012). Other holistic approaches are in general discarded due to lack of time and budget.

As part of a larger research project (Guerberof 2012) that explored correlations between productivity, quality and experience when translating and post-editing machine-translated content, three professional reviewers were recruited to determine quality of the final target texts produced by the 24 translators. Quality was operationalized by counting the number of errors present in the final target texts. With this objective in mind, we engaged three reviewers to verify that the results on quality were consistent and not skewed by either our own evaluation or the assessment of one single reviewer. When the results were analysed, their assessment was not as homogenous as we had, perhaps naively, imagined. In this article the results found after analyzing the data from these reviewers is presented followed by our conclusions on this phase of the project.

Related Work

At the time of writing this article, there were few studies in machine translation and post-editing that had also evaluated the quality of target (post-edited) texts using a group of evaluators. Bowker and

Ehgoetz (2007) explore user acceptance of machine translation output among a group of 121 professors in the Arts Faculty at University of Ottawa. They judge different documents according to speed, quality and cost. The results show that two thirds of the participants (21) chose the post-edited option and one third (10) the human translation. Still, evaluators chose different documents showing that they had different preferences even within a broader category (post-edited versus human translation). The researchers also point out that language professional participants are more linguistically sensitive to language quality than those that are not language professionals, and that the latter might be more tolerant to linguistic errors.

Fiederer and O'Brien (2009) also examine the question of quality in machine-translated texts. They use eleven raters to evaluate sentences according to the Clarity, Accuracy and Style parameters. The raters gave equal scores for translated and post-edited sentences with regards to Clarity, higher scores for post-edited sentences with regards to Accuracy, and finally they gave higher scores to translated sentences in terms of Style. Further, raters chose primarily the translated sentences as their favorite sentences (63 percent of the sentences as opposed to 37 percent for post-edited sentences). The researchers did not show data on inter-rater agreement but it can be seen from the results that raters did not always agree on the options chosen.

Carl et al. (2011) compare the post-editing experience in a group of translation students and professionals. The quality of the translations was evaluated by seven native Danish speakers (four were professional translators). Each evaluator was presented with four translations, two manual translations and two post-edited texts, and they were asked to rank the translations. The results show that there is slight agreement among reviewers when ranking the translations both in terms of inter and intra-coder agreement. Further, this low agreement suggests that the assessment of translation quality is "too difficult" (ibid: 134).

García (2010) explores the use of machine translation and post-editing in a non-professional context. Two markers assessed the quality of the translation and judged the final results. Both markers rated the post-edited segments higher than those translated without MT, but they showed significant differences among them. In 2011, García set up a second phase of the previous study, with 14 students from English to Chinese. Regarding quality, post-edited texts scored higher than translations but evaluators seemed to show a "big disparity" (ibid: 224) in their assessments of the translations.

Koehn (2012) reports on the experience of running evaluation assignments to measure quality of machine translation quality using automatic and manual metrics. He found low agreement with evaluators when assessing fluency and adequacy in output quality, and he pointed to the subjectivity in the judgment of translations made by others and possibly even of translators' own translations.

Outside the realm of post-editing and machine translation, other researchers such as Brunette (2005), Künzli (2006, 2007), and Mossop (2001, 2007a, 2007b) have explored the reviewing task and have pointed to the fact that reviewers might insert preferential changes or even introduce errors during revision.

Material and Methodology

In order to generate the MT output, a Moses (Koehn et al. 2007) engine was trained with a translation memory (TM) and three glossaries. The TM used came from a supply chain management provider (IT domain), and it had 173,255 segments and approximately 1,970,800 words (English source). The Excel glossaries contained 610 entries and 94 entries of core terminology; the XML file contained 9,106 entries (software strings in xml format). The resulting output obtained a BLEU score (Papineni et al. 2002) of 0.60 and a human evaluation score of 4.5 out of 5 points (5 being Excellent and 1 being Very Poor according to predefined criteria).

The file set used in the actual project was a new set of strings from the help system and user interface and therefore different from the parallel data used to train the engine. In order to obtain fuzzy matches, the new files were pre-translated using the option Pre-translate in Trados. The segment pairs were selected together with the corresponding fuzzy match level. The fuzzy match level 85-94% was then used as the Fuzzy match option; and the source segments that fell in the 0-50 percent range were translated using the engine. All duplicates and all segments with a length below 4 and above 26 words were deleted. The three-word segments, without context, can generate a lot of doubts during translation, whereas in the case of segments over 26 words, there were very few in the corpus and there was not sufficient material to replicate them in all categories (No match, Fuzzy match and MT match). All tags with place holders ({ph}) were replaced, and a corpus sampler was applied. The sampler creates a histogram with segment lengths of the original corpus, and it applies the same length distribution to the sample. This same pattern was then applied to each category (No, Fuzzy and MT match) so that the distribution in each category was balanced.

The final file set contained 2,124 words in 149 segments distributed as follows: No match, 749 words, MT match (the output), 757 words, and Fuzzy match, 618 words, from the 85 to 94 percent range.

For this experiment, 24 professional translators from English to Spanish were selected from those approved in a large language service provider (LSP) database. The translators are approved through a selection process that involves CV selection, testing and sample testing within live projects. Since post-editing is a relatively new task for LSPs and the number of post-editors available is limited,

the Vendor Management team usually contacts translators with or without experience when a new post-editing project arrives. Therefore, the same criteria was followed: no post-editing experience was required a priori, but translators with post-editing experience were not discarded.

The 24 translators received the same set of segments, and they had the task to translate the No match and edit the MT and Fuzzy matches (they were not aware of the origin of each proposal). Three initial segments (one of each category) were included for translators to practice and become familiar with the tool, the glossary and the instructions. These were not included when measuring productivity or quality, and they were also not sent to the reviewers.

The criteria for reviewers was, however, different. The three professional reviewers had to have at least three years' experience in localization (software, help and/or documentation) and in Computer Aided Translation Tools (SDL Trados, Déjà Vu, MemoQ or similar tools). Familiarity with tools is an indication of familiarity with translation memories and post-editing, and reviewers needed to be aware of the environment translators were working in to assess the texts produced by them. The reviewers should also had to have at least six months' working experience in MT post-editing and in Business Intelligence software translation. Reviewers needed to have sufficient experience in software translation and in reviewing post-edited material as not to introduce unnecessary changes and, at the same time, perform the assignment at the standard review speed.

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

The final texts from the translators were evaluated using the LISA QA model. The reviewers were informed that they would have 24 individual translations in Word (24 translated versions of the same source text), 24 LISA forms in Excel to complete, and a timesheet in Excel in which to enter the time invested in the review task. They were also informed that they would need to review each translation in Word using the Track Changes option and then record the number of errors per category in the LISA form, one per translator.

Each Word document contained seven columns: Segment ID, Source Segment, Target Segment, Type of Match, Post-edited target, PE difference in %, and Type of error. The Target Segment could be blank if translators were not given a TM or MT proposal, and they had to fully translate the English segment into Spanish. The reviewers were instructed that if the same error occurred more than once in one translation, it should be counted only once. The reviewers were informed about the

Type of Match (Fuzzy, MT or No match) because they were requested to mark any overcorrection found, although they were not to consider them as errors.

To track the time, the reviewers were given an Excel timesheet. They were told that they could go back and add more time if they realized, after completing the first review, that they wanted to change a correction or correct the text even further. They were also informed that it would be logical to have different times invested per translator because they were correcting the same text repeatedly. Since real-time data was needed, it was not necessary to change the times to make them consistent for all texts. In the case of the reviewers, there was no possibility to time them automatically because of the reviewing method applied.

There was a separate section to explain the review process in itself. The reviewers had to read the source text, then the proposed TM and MT translation, and the final target text. They were also advised not to insert any preferential change (to only correct errors), to make minimal numbers of changes, to make sure the translators followed the glossary provided, and to be consistent across the 24 translations. Finally, they were given some clarifications on error typology according to the LISA QA model. They were also asked to be flexible as their ability to spot errors was not being judged, but the quality of the final translation. There were separate instructions on how to use the LISA form. Finally, they were given the same quality expectations, style and terminology instructions as the translators.

LISA QA process

LISA defines different types of error: Mistranslation, Accuracy, Terminology, Language, Style, Country, Consistency and Format. Mistranslation refers to the incorrect understanding of the source text; Accuracy to omissions, additions, cross-references, headers and footers and not reflecting the source text properly; Terminology to glossary adherence; Language to grammar, semantics, spelling, punctuation; Style to adherence to style guides; Country to country standards and local suitability; Consistency to coherence in terminology across the project; and Format to correct use of tags, correct character styles, correct footnotes translation, hotkeys not duplicated, correct flagging, correct resizing, correct use of parser, template or project settings file.

The errors found are then assigned a severity level: Minor, Major or Critical. All errors are weighted according to these categories. For example, an error classified as Minor carries one point, if classified as Major, five points, and finally if it is deemed to be Critical it is penalized with the total amount of allowed points plus one (that is if a translation is allowed to have 5 points according to the volume, and there is one error with a Critical severity level, then it will have a value of 6 points).

The text might Pass or Fail the quality metric if the number of errors exceeds the points allowed for that particular number of words. Our focus was on the number and classification of errors, as the scope of this study was not to establish if a particular translator offered a good or poor performance (Pass or Fail), but whether the number and type of errors were affected by the use of a translation tool and therefore if the errors had an impact on the overall productivity of the translation. In other words, we needed to establish whether the time saved using MT or TM meant additional time to fix errors at a later stage in the localization process.

Three categories were introduced under the type of error: No match, MT match or Fuzzy match segments. This was so that reviewers could insert each error in the category where the error was found.

Results

The reviewers sent back 24 LISA forms, 24 edited Word documents (with tracked changes) and one timesheet with the registered time employed to correct each text and to complete each form. With this information in hand, all errors were transferred to three databases for statistical analysis.

Results on reviewers' time

Figure 1 shows the aggregated time reviewers took to complete the task, that is, the time to correct each translator is plotted according to the results from each reviewer. The y-axis shows the time per test from 0 hours to 3 hours. Each reviewer had 24 translations and that meant a total of 50,976 words to review. They were dealing, however, with one text repeated 24 times. When queried, the reviewers confirmed that they had reviewed in strict order from 1 to 24, although on occasions they had to go back to change some corrections in previous translations.

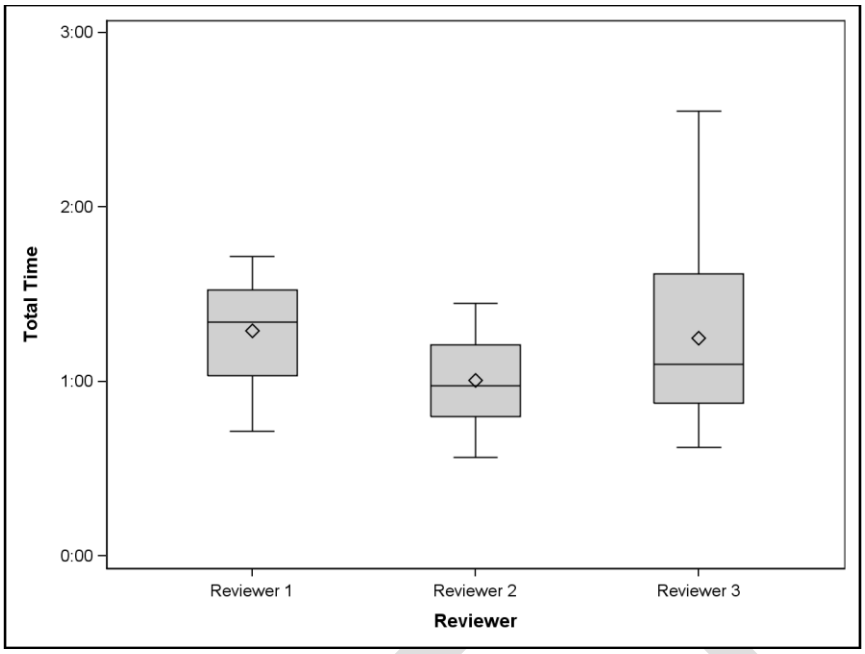

Figure 1: Total reviewing time

Figure 1 shows that times are different for the three reviewers. The time Reviewers 1 and 3 spent in the task are similar (represented by the size of the boxes containing the data from the first to the third quartile); their approaches to the tasks, however, seem to be different because the mean and median values (indicated by the diamond and the black line inside the box respectively) are different, and Reviewer 3 presents a wider range of reviewing time. This reviewer took longer initially than the other two (more than 2 hours for one translator as it can be seen in the upper whisker), and then she gained speed as the text became familiar. In the end, however, she corrected certain texts more quickly than Reviewer 1 (this can be appreciated in the lower whisker for Reviewer 3, which is below 1 hour). Although Reviewer 2 was faster, she forgot to include the types of error in the Word files, and she had to go back and correct the mistake, adding the time to the Excel form. It is possible that the time keeping was somewhat distorted as a result, but it could also be that she was simply faster at correcting all the texts. Table 1 shows these same values numerically to appreciate the information in detail.

Reviewer	N	Min	Quartile 1	Mean	Median	Quartile 3	Max	SD	Range	Q Range
Rev 1	24	0:43:00	1:02:00	1:17:25	1:20:30	1:31:30	1:43:00	0:16:38	1:00:00	0:29:30
Rev 2	24	0:34:00	0:48:00	1:00:25	0:58:30	1:12:30	1:27:00	0:15:25	0:53:00	0:24:30

Reviewer	N	Min	Quartile 1	Mean	Median	Quartile 3	Max	SD	Range	Q Range
Rev 3	24	0:37:20	0:52:30	1:14:58	1:06:00	1:37:00	2:33:00	0:29:29	1:55:40	0:44:30

Table 1: Descriptive data of review time

Reviewer 2 was the fastest reviewer if the mean value is considered, but also if the minimum and maximum values are considered. This means that her longest review took 1 hour and 27 minutes and the shortest was 34 minutes. Reviewer 1 has a higher mean value than Reviewer 3 but he was more consistent regarding time over the 24 translations, since the minimum value is 43 minutes and the maximum 1 hour and 43 minutes. Reviewer 3 has a wider range (44:30) because she took a maximum of 2 hours and 33 minutes (for the first test) and a minimum of 37 minutes and 20 seconds. Therefore, Reviewer 2 is the faster overall and also per individual test. Reviewer 1 is the slowest overall but more consistent with each individual text in terms of timing. Figure 2 and Table 2 show the number of words corrected per minute. Figure 1 and Table 1 show the number of words corrected per minute.

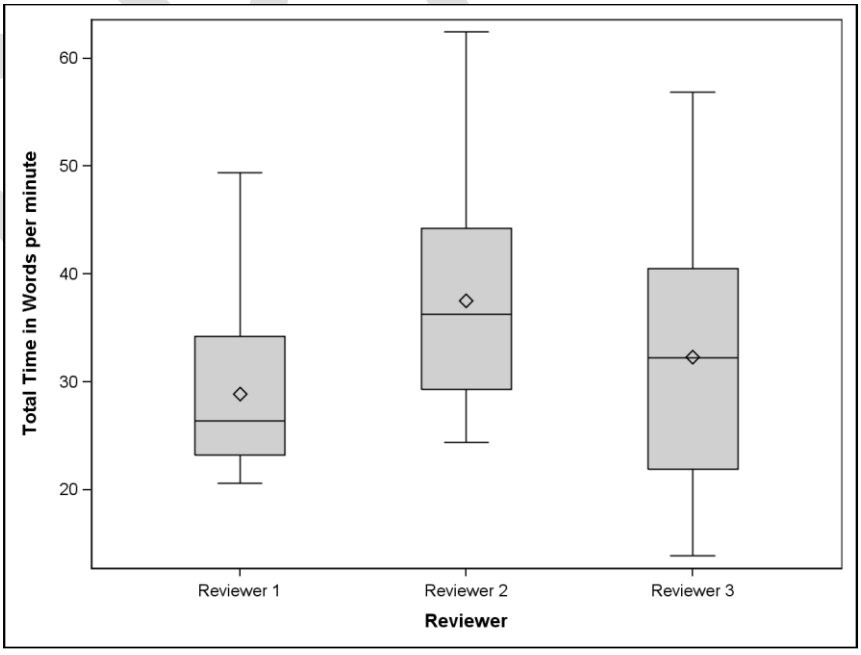

Figure 2: Total reviewing time in words per minute (WPM)

Reviewer	N	Min	Q1	Median	Mean	Q3	Max	SD	Range	QRange
Rev 1	24	20.62	23.21	26.39	28.91	34.26	49.40	7.36	28.77	11.04
Rev 2	24	24.41	29.31	36.31	37.50	44.25	62.47	9.94	38.06	14.94
Rev 3	24	13.88	21.90	32.25	32.27	40.49	56.89	11.24	43.01	18.59

Table 2: Descriptive data of review time in WPM

Reviewer 2 has the fastest reviewing time (more than 62.47 words per minute), followed by 3 and (56.89 wpm) then by 1 (49.40 wpm). Reviewer 1 seems to review at a more similar pace throughout the test (the standard deviation is 7.36) and Reviewer 3, with a deviation of 11.24, has a wider range of 18.59 words per minute. If the data is modelled with repeated measures (Translator) taking *logarithm of Total time in words per minute* as the response variable and *Reviewer* as the explanatory variable, there are statistically significant differences among the three reviewers ($F=17.30$; $p<0.0001$). However, there are no statistically significant differences between Reviewers 1 and 3 with regards to time.

Results on reviewers' errors

In this section, the data is analysed to explore if the three reviewers agreed on the number of errors found. With this objective in mind, an error indicator is defined for those segments that were highlighted by reviewers as containing an error:

- 0 means that there are no errors
- 1 means that there are errors

To examine the degree of agreement among reviewers, another variable is defined with the following values:

- There are no differences among translators

- Reviewer 1 does not agree with Reviewers 2 and 3
- Reviewer 2 does not agree with Reviewers 1 and 3
- Reviewer 3 does not agree with Reviewers 1 and 2.

The error indicator will only highlight whether those segments contain an error or not, not the number of errors per segment. However, it is important to note that, out of the 10,728 segments, the reviewers marked mostly 1 error per segment; only in two segments did they mark 3 errors, and in 12 segments out of 10,728 they marked 2 errors (10,728 segments if we consider that there were 149 segments times 24 translators' times three reviewers). This should give us an idea of the degree of agreement on the number of segments in which corrections had to be made.

Reviewers' agreement	**Fuzzy match**		**MT match**		**No match**		**All**	
	N	**%**	**N**	**%**	**N**	**%**	**N**	**%**
No differences	953	79.42	866	73.64	794	66.17	2613	73.07
Reviewer 1 does not agree	64	5.33	109	9.27	142	11.83	315	8.81
Reviewer 2 does not agree	103	8.58	103	8.76	137	11.42	343	9.59
Reviewer 3 does not agree	80	6.67	98	8.33	127	10.58	305	8.53

Table 3: Percentage of error indicator

For the 24 translators there are a total of 3,576 segments (149 segments times 24 translators). The column All shows that the reviewers agree on 73.07 percent of all segments (2,613), and they disagree on 26.93 percent (963 segments). There is more agreement on the Fuzzy matches (79.42 percent) and less on the No matches (66.17 percent). The data above indicate solely in which segments an error was marked; it does not tell if there was agreement on the type of errors marked.

Reviewer	**Fuzzy match**		**MT match**		**No match**		**All**	
	N	**Error #**	**N**	**Error #**	**N**	**Error #**	**N**	**Error #**
Reviewer 1	1200	187	1176	171	1200	309	3576	667

Reviewer	Fuzzy match		MT match		No match		All	
	N	Error #	N	Error #	N	Error #	N	Error #
Reviewer 2	1200	206	1176	173	1200	287	3576	666
Reviewer 3	1200	149	1176	232	1200	352	3576	733

Table 4: Total number of errors

Table 4 shows the absolute numbers of segments containing errors. The number of words is not considered and, since it was slightly different per category, the ratio of errors per word might differ. Reviewers 1 and 2 differ in the total number of segments containing errors by only one, yet these are distributed differently across two categories (No match and Fuzzy match in particular), and they show very similar results in the MT match category, a difference of only two errors. Reviewer 3 shows a higher number of errors than the other two reviewers in all categories but Fuzzy matches, although they all agree that the No match category has more errors. The number of segments containing errors per translator and category is also different between the three reviewers.

Rev	Time /Errors	N	Min	Q1	Mean	Median	Q3	Max	SD	QRange	Range
Rev 1	Total Time	24	20.62	23.21	28.91	26.39	34.26	49.40	7.36	11.04	28.77
	Total Errors	24	13.00	18.00	27.79	23.50	32.00	60.00	12.89	14.00	47.00
Rev 2	Total Time	24	24.41	29.31	37.50	36.31	44.25	62.47	9.94	14.94	38.06
	Total Errors	24	14.00	19.00	27.75	24.00	35.00	57.00	11.97	16.00	43.00
Rev 3	Total Time	24	13.88	21.90	32.27	32.25	40.49	56.89	11.24	18.59	43.01
	Total Errors	24	12.00	19.50	30.71	23.50	38.50	64.00	15.19	19.00	52.00

Table 5: Descriptive data on errors per reviewer

By simply looking at these descriptive data in Table 5 from the reviewers it can be seen that the minimum and maximum values are quite similar. In other words, the reviewers agree on the very poor and very good results. The mean value for Reviewer 3 is slightly higher: she made more corrections overall. For Quartile 1 there is relative agreement, suggesting that it might be easier to agree on the translators that made minimum and median errors (Minimum, Quartile 1 and Median values), than those that made more errors (Quartile 3 and Maximum values) possibly because once there are more errors, one reviewer might decide to correct more things to his or her taste.

To see if there are statistically significant differences, a linear regression model with repeated measures taking the *logarithm of Words per minute* (in this case, words per minute to review) as the response variable and *Total errors* and *Reviewer* as explanatory variables. There are statistically significant differences between the Reviewers ($F=18.00$; $p<0.0001$) and for the *Total errors* ($F=14.39$; $p=0.0004$).

Figure 3 and Figure 4 illustrate these findings, the former figure highlights the time lines, and the latter the error lines for the three reviewers.

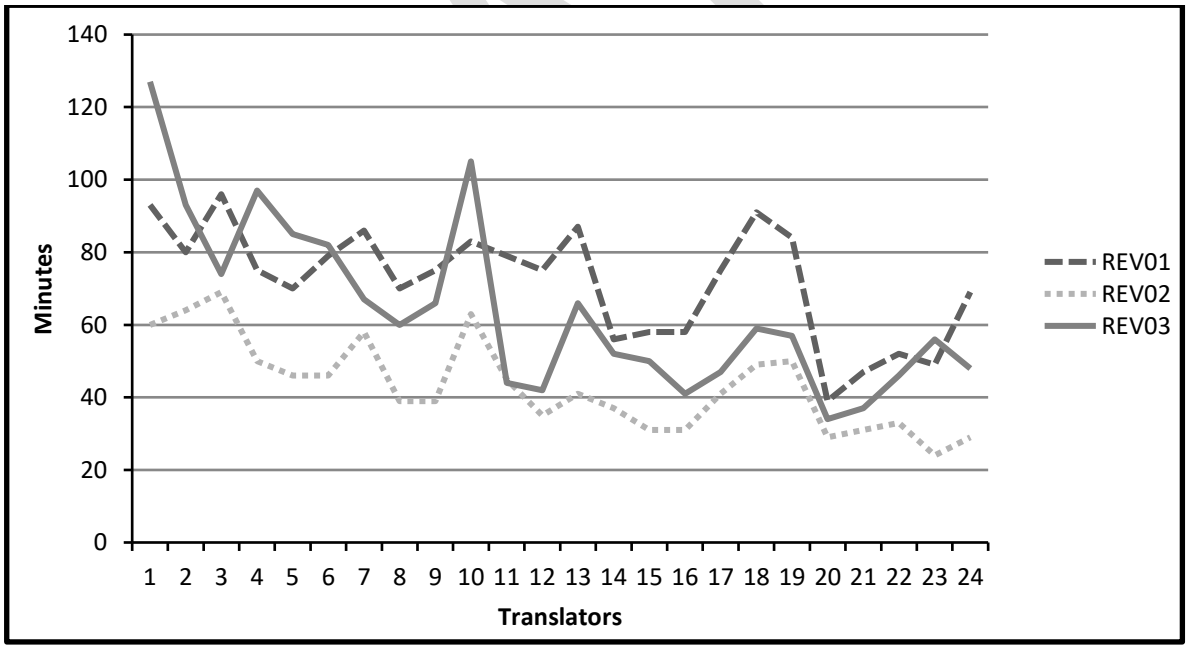

Figure 3: Total time in minutes taken by each Reviewer

Figure 3 shows how reviewers gained speed, investing fewer minutes, as the 24 texts were reviewed. However, in particular cases, the speed decreases for those translators that have more errors, even at the latest stage of the review (Translators 10, 13, 18, 19, 22 and 24).

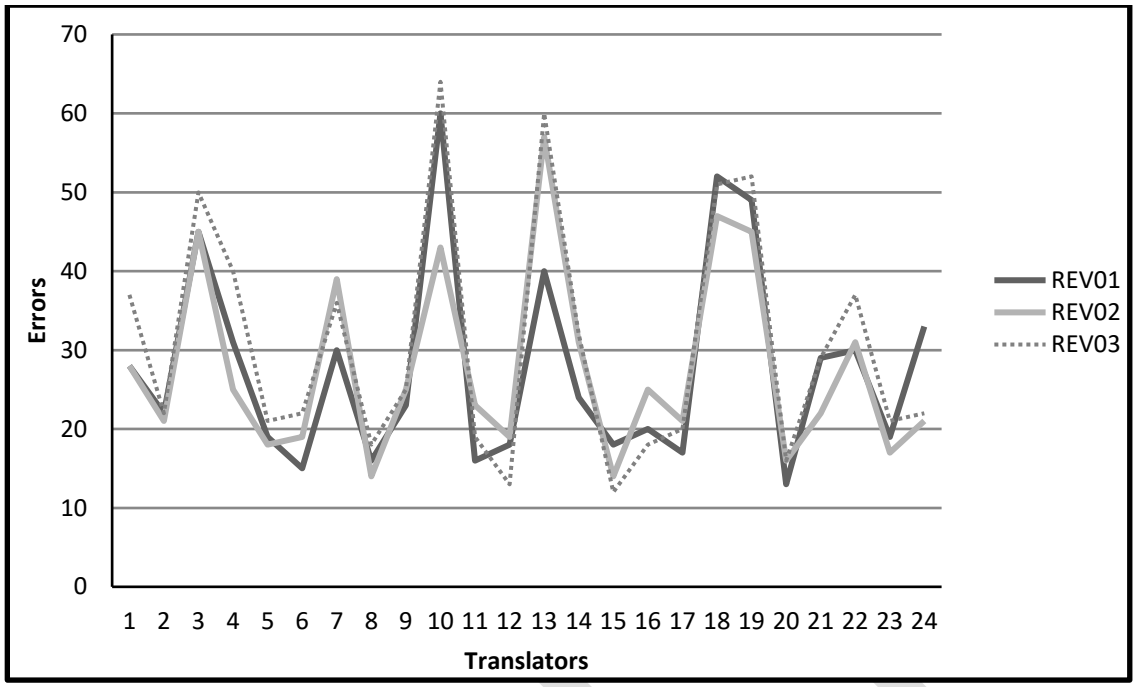

Figure 4: Total errors reviewers

In Figure 4, the speed is not directly related to the number of errors found. There are more errors for Translators 3, 7, 10, 13, 18, 19 and 22 (peaks) and fewer for Translators 2, 5, 6, 8, 12, 15, 16, 17 and 20. The three lines representing the errors marked by the reviewers, however, are not exactly the same per translator.

There is no agreement between the three reviewers, but we have seen that Reviewer 1 and Reviewer 2 had a similar number of global errors. It is important to remember that the data analyzed has a hierarchical structure (that is, there are 24 translators, 149 segments per translator, three Match categories, 8 Types of errors, and three reviewers), the fact that the global number of errors (between Reviewers 1 and 2, for example) is similar does not necessarily mean that the reviewers agree on the errors marked for each translator in each segment for each category.

Comparing reviewers

In order to further explore the agreement or association between the three reviewers in terms of numbers of mistakes, errors were classified in three categories: Few errors, Average errors, and Many errors. Since the number of errors that the reviewers marked in the three translation categories (Fuzzy, MT and No match) was different, as we have seen above, the first and third quartiles were used to determine where to divide the data. The results were:

For Fuzzy and MT match:

- Between 0 and 5 errors;

- Between 6 and 10 errors;
- More than 10 errors.

For No match:

- Between 0 and 8 errors;
- Between 9 and 16 errors;
- More than 16 errors.

The contingency tables were calculated and a Kappa coefficient applied per Match category and Reviewer. In statistics, the Kappa coefficient is a measurement to indicate inter-rater or inter-annotator agreement. For the No match category there is less variability, the Kappa coefficient for the categorization of number of errors is from 0.61 to 0.75 (p=<.0001), and this means that there is agreement between the three reviewers. On the other hand, for Fuzzy and MT match, the Kappa measure is more variable, and it is between values that show either no agreement (-0.02 ≈ 0) or weak agreement (0.43). The only cases where the relation (disagreement in this case) is clear are:

- For Fuzzy match, Reviewer 2 vs. Reviewer 3: There is no agreement between these two reviewers.
- For MT match, Reviewer 2 vs. Reviewer 3: There is no agreement between these two reviewers.

Table 6 illustrates the agreements or associations between the three reviewers according to the Match categories.

Match	Contingency table	Kappa	95% Lower	95% Upper	Two-sided Pr > |Z|
Fuzzy	Reviewer_1 * Reviewer_2	0.43	0.14	0.72	0.00
Fuzzy	Reviewer_1 * Reviewer_3	0.30	0.02	0.59	0.02
Fuzzy	Reviewer_2 * Reviewer_3	-0.02	-0.24	0.20	0.84
MT	Reviewer_1 * Reviewer_2	0.36	0.05	0.68	0.01
MT	Reviewer_1 * Reviewer_3	0.39	0.12	0.66	0.00
MT	Reviewer_2 * Reviewer_3	0.17	-0.12	0.46	0.21

No match	Reviewer_1 * Reviewer_2	0.62	0.37	0.87	<.0001
No match	Reviewer_1 * Reviewer_3	0.61	0.34	0.88	<.0001
No match	Reviewer_2 * Reviewer_3	0.75	0.53	0.97	<.0001

Table 6: Kappa statistical values according to Match category and Reviewer

Therefore, the three reviewers agree on the No match category and Reviewer 1 slightly agrees with Reviewers 2 and 3 in Fuzzy and MT match but Reviewers 2 and 3 do not agree. Perhaps the fact that the reviewers knew the origin of the segments, whether Fuzzy or MT, influenced them when marking the errors. It is difficult to say. In this analysis, we are comparing them in terms of errors per translator, but still they might disagree per segment or per classification of errors. Below we present one samples taken from the actual reviewed files to illustrate this divergence in the reviewing criteria. The target translations are taken from one translator as an example and the corrections made by each reviewer in the first three segments are shown.

Reviewer 1

Segment·ID¤	Source·Segment¤	MT·Target·Segment¤	Post-edited·target¤	TM·Match¤	PE·difference(%)¤	Type·of·error¤
4¤	The·name·of·the·shortcut·is·updated.¤	El·nombre·del·objeto·se·actualizará.¤	El·nombre·del·acceso·directo·se·actualizará.¤	Fuzzy·match¤	82,93¤	¤
5¤	You·can·filter·the·data·displayed·in·the·analysis·by·selecting·options·from·the·Filters·panel.¤	¤	Para~~ude~~·filtrar·los·datos·presentados·~~por~~·en·el·análisis,·seleccion~~ando~~·las·opciones·del·panel·Filtros.¤	No·match¤	0¤	Language·(1)¤
6¤	This·procedure·assumes·that·you·have·already·created·an·analysis·and·added·at·least·two·attributes·to·the·Filters·panel.¤	En·este·procedimiento·se·da·por·hecho·que·ya·se·ha·creado·un·análisis·y·ha·agregado·al·menos·dos·atributos·al·panel·de·filtros.¤	En·este·procedimiento·se·da·por·hecho·que·ya·se·ha·creado·un·análisis·y·que·se·han·agregado·al·menos·dos·atributos·al·panel·Filtros.¤	MT·match¤	95,02¤	¤

Figure 5: Sample 1 correction Reviewer 1

In Figure 5 Reviewer 1 only corrected segment 5 as a Language error. The problem appears to be that this translator used a *gerundio* (Spanish gerund) to construct the sentence (No match) and the reviewer found that this was a mistake. Although the *gerundio modal* (used in this sample) is accepted in Spanish (see Real Academia Española 2009 and Fundeu 2012), the *gerundio de consecuencia* in English is wrongly translated in Spanish as a *gerundio*. Therefore, this might be the reason why Reviewer 1 corrected it.

Reviewer 2 on the other hand did not correct any of the first 3 segments. She deemed the translations correct. Finally, Reviewer 3 highlighted the same error in segment 5 as Reviewer 1 did and also another error in segment 4. The phrase "is updated" was translated as "*se actualizará*" (future) instead of "*se actualiza*" (present), and this was deemed to be an Accuracy error (in the Fuzzy match category). In English, the difference would be between "is updated" and "will be updated". Reviewers 1 and 2 did not correct this, possibly because the present tense here does not necessarily indicate that the action has finished, and therefore both translations are potentially correct. All reviewers agreed that segment 6 did not require any change: Reviewers 1 and 3 agreed that segment 5 required the same change; Reviewers 1 and 2 agreed that segment 4 did not require any change, but Reviewer 3 disagreed.

There are many more examples of these agreements and disagreements. We have picked one that represents the type of corrections and disagreements in all 24 texts. On the one hand, it is understandable that by having three reviewers there is an exposure to three different versions, even though great emphasis was placed on not introducing any preferential changes. On the other hand, we were quite surprised at the disagreements and at some of the changes made by the three reviewers.

Error classification

We were also interested in seeing the error behaviour in terms of type of errors and Match category. Errors have been analyzed and distributed according to the LISA standard to see if the typology of errors varies depending on the type of proposed text (Fuzzy or MT match) or without any translation proposal. Since there are three reviewers in this case, let us look at the similarities in categorization of the errors.

Type of error	**No match**	**MT match**	**Fuzzy match**	**Totals**	**% No match**	**% MT match**	**% Fuzzy match**	**% Total**
Mistranslation	0	32	14	46	0%	70%	30%	**7%**
Accuracy	30	11	45	86	35%	13%	52%	**13%**
Terminology	89	32	73	194	46%	16%	38%	**29%**
Language	124	64	43	231	54%	28%	19%	**35%**
Consistency	1	0	1	2	50%	0%	50%	**0%**
Country	0	0	0	0	0%	0%	0%	**0%**
Format	3	11	3	17	18%	65%	18%	**3%**

Style	53	21	8	82	65%	26%	10%	**12%**
Totals	**309**	**171**	**187**	**667**	**46%**	**26%**	**28%**	**100%**

Table 7: Reviewer 1 number and percent of errors per type of error

Table 7 shows results for Reviewer 1: 46 percent of all errors are No match. The total results for Fuzzy and MT are 28 and 26 percent respectively. Regarding categories, No match has the most errors in the categories Language, Style and Terminology. MT has the most in the categories Mistranslation and Format. Fuzzy match has the most in the Accuracy category. Terminology is low in MT and Style is low in Fuzzy matches as well.

In the case of reviewer 2, the No match has the highest number of errors with 43 percent, followed by Fuzzy matches with 31 percent and lastly by MT matches with 26 percent. The highest number of Language, Style and Accuracy errors are placed within the No match category, while in this case Terminology and Country errors are the highest number in the Fuzzy matches. MT still has the highest number of Mistranslation errors, albeit the number of this type of errors recorded by Reviewer 2 is very low. Style errors are low in Fuzzy matches and Terminology in MT matches.

Finally, Reviewer 3, places the majority of errors in the No match category (48 percent), followed by MT matches with 32 and lastly Fuzzy matches with 20 percent. No match predominates in Style, Language, Terminology and Format. Fuzzy match has the highest number of errors in Accuracy, followed by Terminology, and Format and MT match in Mistranslation, followed by Style and Format.

Although there are differences in the classification of errors by the reviewers, they all agree that the No match category has more errors overall and that Language and Style were problematic areas in this particular category. In the case of Fuzzy matches, Accuracy seems to present problems, and for MT matches, Mistranslations. This seems quite logical: in the case of No matches, these strings have never been reviewed (this was, in fact, the first time they were translated) while in the case of translation memories and machine translation, the segments had been "extracted" from the original translation memory. In the case of Fuzzy matches, the main changes to be made in the 85-94% range are related to single words and therefore if this change is missed, there will be a greater probability of Accuracy errors. In the case of MT, the engine might produce an output that is different in meaning, which would need to be completely rearranged or rewritten, thus causing Mistranslation errors. Overall, however, the No match category is not exempt from this type of error. Another interesting aspect is that Fuzzy match has a low percentage of Style errors, indicating the high quality of

the original TM, and MT has a low percentage of Terminology errors. We are unsure about the reasons for this, since the same TM was used to train the engine. It could be that translators when correcting blatant errors in MT segments consulted the glossary more frequently.

Overcorrections

The reviewers were instructed to mark overcorrections, that is, to mark the edits or post-edits that translators had made but that went beyond what was needed, and, at the same time, they were instructed not to count them as errors.

Reviewers 1 and 2 marked very few overcorrections, 4 and 10 respectively while Reviewer 3 marked 146 in total. This might reflect the fact that the reviewers were informed that overcorrections were not to be marked as errors, hence, Reviewers 1 and 2 did not think they were important, while Reviewer 3 spent more time marking this (this Reviewer also identified more errors overall). Perhaps, the instructions should have been clearer on what an overcorrection was and how to classify it. The only conclusion we can draw from the figures is that Reviewers 1 and 3 found more overcorrections in Fuzzy match than in MT match, although Reviewer 1 found very few overall, and Reviewer 2 found an equal amount of corrections in both categories. Although the aim of this study was not to investigate the concept of preferential changes, we believe that this is a topic that would need to be studied further in order to develop instructions or training materials for post-editing.

Conclusions

According to the data received from the three reviewers, they agreed that most segments did not contain errors and that the No match category had a higher percentage of segments with errors overall. The percentage of segments with errors in Fuzzy and MT match had almost identical values, 18.17 and 18.11 percent respectively. However, we also observed that the reviewers behaved differently when correcting the translations from the 24 translators both in time, words reviewed per minute, and in number of errors.

When comparing the reviewers against each other, we have found that they tended to agree on the general number of errors found in the No match category but that their agreement in Fuzzy and MT match was either weak or there was no agreement, perhaps indicating that the origin of the text might have influenced their evaluation. The reviewers also tended to agree on best and worst performers in general, but there was great disparity in the translators' classifications if they were ranked according to the number of errors. These disagreements could also mean that each reviewer might

adapt the instructions to their own particular logic if grey areas are perceived, or that each reviewer focuses on areas of particular interest (for example, certain grammatical errors) that they have established from their previous experience, or that the source text can be interpreted in different ways, especially in the absence of context.

Regarding the type of errors found in the study, and although there are differences in the classification of errors according to each of the reviewers, they all agree that the No match category has more errors overall and that Language and Style were problematic areas in this particular category. In the case of Fuzzy matches, Accuracy seems to present problems, and for MT matches, Mistranslations. However, Terminology errors are low in MT matches and Style errors are low in Fuzzy matches. With regards to overcorrections, Reviewers 1 and 2 found very few overall, and Reviewer 3 found significantly more in some translators. Reviewers 1 and 3 found more overcorrections in Fuzzy match than in MT match, and Reviewer 2 found an equal amount of corrections in both categories.

Therefore, we can say that although in general terms reviewers might agree on blatant errors and translators that perform poorly or quite well, they do not always agree on the marking and classification of errors, and that they might show significant differences if we look at timing and number of errors per categories.

Due to the importance that reviewing texts and marking errors have in the new localization setups where MT is increasingly present, we feel that more research need to focus in identifying the agreements and disagreements among reviewers and elaborating possible reasons for these differences. If we know the areas were there are disagreements and the reasons behind these disagreements, the design of reviewing instructions or reviewing training can improve so that the reviewer adds more value to the final text.

References

Bowker, L. Ehgoetz, M. 2007. "Exploring User Acceptance of Machine Translation Output: A Recipient Evaluation". Kenny, D. and Ryou, K., eds. 2007. *Across Boundaries: International Perspectives on Translation*. Newcastle-upon-Tyne: Cambridge Scholars Publishing. 209-224

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

Carl, M. Dragsted, B. Elming, J. Hardt, D. Jakobsen, A. 2011. "The process of post-editing: a pilot study". In *Proceedings of the 8th international NLPSC workshop.* Bernadette Sharp, Michael

Zock, Michael Carl, Arnt Lykke Jakobsen (eds). (Copenhagen Studies in Language 41), Frederiksberg: Samfundslitteratur: 131-142. Available from http://www.mt-archive.info/NLPCS-2011-Carl-1.pdf

Fiederer, R. O'Brien, S. 2009. "Quality and machine translation: a realistic objective?" *The Journal of Specialised Translation*. (11). Available from http://www.jostrans.org/issue11/art_fiederer_obrien.pdf Accessed June 2012.

Fundeu. 2012. Wikilengua. http://www.wikilengua.org/index.php/Gerundio

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

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

Guerberof, A. 2012. Productivity and quality in the post-editing of outputs from translation memories and machine translation. PhD Thesis. Universitat Roviera I Virgili. Available from http://www.tdx.cat/handle/10803/90247. Accessed June 2014.

Koehn, P. 2012. "What is a Better Translation?" Reflections on Six Years of Running Evaluation Campaigns". *Tralogy* [En ligne session 5- Quality in Translation / La qualité en traduction, mis à jour le 31/01/2012] Available from http://homepages.inf.ed.ac.uk/pkoehn/publications/tralogy11.pdf

Koehn, P. Hoang, H. Birch, A. Callison-Burch, C. Federico, M. Bertoldi, N. Cowan, B. Shen, W. Moran, C. Zens, R. Dyer, C. J. Bojar, O. Constantin, A. and Herbst, E. 2007. "Moses: Open source toolkit for statistical machine translation". In *Proceedings of the 45th Annual Meeting of the ACL Companion Volume Proceedings of the Demo and Poster Sessions.* Association for Computational Linguistics. Prague: 177–180.

Künzli, A. 2006. "Translation revision - A study of the performance of ten professional translators revising a technical text". *Insights into specialized translation*. Maurizio Gotti and Susan Sarcevic (eds). Bern & Frankfurt: Peter Lang: 195-214.

Künzli, A. 2007. "Translation Revision. A study of the performance of ten professional translators revising a legal text". *Doubts and Directions in Translations Studies. Selected contributions from EST Congress Lisbon 2004*. Gambier, Y. Shlesinger, M. Stotlze, R. (eds.). Amsterdam and Philadelphia: Benjamins. 115-126.

Mossop, B. 2001. 2007a. *Editing and Revising for Translators*. Manchester: St. Jerome Publishing.

Mossop, B. 2007b. "Empirical studies of revision: what we know and need to know". *Journal of Specialised Translation*. Issue 8. Available from http://www.jostrans.org/issue08/art_mossop.pdf. Accessed June 2012.

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

Papineni, K. Roukos, S. Ward, T. Zhu, W.J. 2002. "BLEU: A method for automatic evaluation of machine translation". In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*. Philadelphia: 311-318. Also available from http://acl.ldc.upenn.edu/P/P02/P02-1040.pdf. Accessed June 2012.

Real Academia Española. Asociación de Academias Americanas. 2009. *Nueva Gramática de la Lengua Española*. Ignacio Bosque, ed. Madrid: Espasa Libros, S.L.U.