Statistical & Neural MT Systems in the Motorcycling Domain for Less Frequent Language Pairs - How Do Professional Post-editors Perform?

Clara Ginovart Cid
Pompeu Fabra University, Barcelona, Spain
Datawords Datasia, Levallois-Perret, France
claratranslator@gmail.com

Abstract

As more language service providers (LSP) are including post-editing (PE) of machine translation (MT) in their workflow, we see how studies on quality evaluation of MT output become more and more important. We report findings from a user study that evaluates three MT engines (two phrase-based and one neural) from French into Spanish and Italian. We describe results from two text types: product description and blog post, both from a motorcycling website that was actually translated by Datawords Datasia. We use task-based evaluation (PE is the task), automatic evaluation metrics (BLEU, edit distance, and HTER) and human evaluation through ranking to establish which system requires less PE effort and we set the basis for a method to decide when an LSP could use MT and how to evaluate the output. Unfortunately, large parallel corpora are unavailable for some language pairs and domains. Motorcycling and the French language are low-resourced, and this represents the main limitation to this user study. It especially affects the performance of the neural model.

Keywords: NMT, post-editing, quality evaluation, machine translation

1 Introduction

As stated by Lommel and DePalma (2016), in Europe, post-editing of machine translation (PEMT) is offered by over 56% of LSPs, and translation volumes will increase by 67% over current [2016] levels by 2020. Thanks to the publication of ISO 18587:2017, we know now that the TSP (Translation Service Provider) is responsible for meeting the quality requirements in a PEMT project and any other specifications agreed in a client-TSP agreement. However, we claim that the post-editor should be responsible for meeting the specifications in the PE assignment according to the TSP-post-editor agreement, which is for the moment not mentioned in the PEMT literature.

For the present article, we have mainly based our methodology on previous studies, such as Béchara et al. (2017), Castilho et al. (2017), Forcada et al. (2017), Isabelle et al. (2017), Van Der Meer et al. (2017), Way (2018), but also on O’Brien (2011), who focuses on correlations between two automatic metrics for MT quality evaluation.

In this user study, a real scenario is used by selecting two translation jobs from Datawords where PEMT is applied and we perform both human and automatic evaluation to compare technical and temporal effort to determine:
- factors to be considered when deciding if it is worth carrying out a translation job with PEMT;
- which MT system (phrase-based MT -PBMT- or neural MT -NMT-) performs better in an in domain setting for French (FR) into Spanish (ES), and French into Italian (IT).

In 2 Method, we explain the nature of the experiment and the technology and process used to select the sample, clean the data, select the participants, and train the MT engines. In
3 Analysis, we explore the reports of each participant, and we discuss the overall results in section 4 Results. Finally, some improvements and new ideas for an upcoming research project are suggested in 5 Further Work.

2 Method

This user study focuses on the comparison of three models: two PBMT (SDL Language Cloud\(^1\) and KantanMT\(^2\)) and one neural (KantanMT). Two different text types were post-edited: one product description (repetitive, 76 segments with an average of 13.25 words per segment) and a blog post (creative, 57 segments with an average of 16.84 words per segment). Together a total of 133 translation units (TU) with 1,967 source words, to be post-edited by a pool of 10 post-editors per language. The two files were also translated with Google Translate (GT) for the purpose of comparing the automatic scores of HTER metric of our engines to a general one.

For both files every segment appeared 3 times in the source column and had 3 different MT outputs in the target column (which amounts to 5,901 source words). SDL, KantanMT Stat, and KantanMT NMT were alternated\(^3\) to avoid any preference by the post-editors regarding the order of appearance in Trados Studio. Therefore, Auto-Propagation was disabled for the post-editing task.\(^4\) Furthermore, every fourth segment contained a question\(^5\) where the participant communicated preference for one MT system output over the other two with a numeric answer.

The participants were required to use SDL Trados Studio as post-editing tool; Flashback Recorder to record their screens during the experiment\(^6\) and Qualitivity\(^7\) plug-in to produce Excel and XML reports of time spent per segment, edit distance\(^8\), and post-editing modification; referred as “PEM%”. The formula is shown below:

\[
100 – (\text{edit distance} / \text{relative edit distance}) \times 100
\]

“Relative edit distance” is the number of characters of the longest segment, be it original target or updated target (Andrew, 2018). What we obtained is a similarity rate. [The] recalculated edit distance is included with the reports that are generated by Qualitivity but not maintained physically with the records in the database and/or currently exported with the raw data.\(^9\) Indeed, in the HTML report both the PEM% and edit distance were recalculated when a segment was revisited as shown in the left part of Figure 1.

We consider that the fact that a post-editor keeps or discards a first change should not disrupt the overall post-editing effort as understood by any TSP-post-editor agreement, and therefore, we calculated the total of additions, deletions and shifts on the XML or Excel report. However, we also calculated the average PEM% between both records and, since it is a similarity\(^10\) percentage and we are interested in the difference between the original target

---

\(^1\) https://languagecloud.sdl.com/translation-toolkit/login

\(^2\) https://app.kantanmt.com/index.php

\(^3\) The first TU showed the translations ABC, where A is SDL, B is KantanMT Stat and C is KantanMT NMT. The second, BCA. And the third, CAB. And so on and so forth.

\(^4\) This guideline is sent to participants in the Instructions sheet. A warm-up activity is performed some weeks before the experiment to confirm that the instructions are understood, and that the software required work well.

\(^5\) Merci d’indiquer quelle traduction vous préférez avec : « 1 », « 2 » ou « 3 », si aucune, rentrez « 0 ».

\(^6\) Except for one Italian participant, due to confidentiality reasons.

\(^7\) Source: https://tinyurl.com/qualitivity

\(^8\) Defined as the minimum edit distance between a translation output and targeted reference translation created by post-editing the output. Edit distance as calculated by Qualitivity plug-in (https://tinyurl.com/DLedist).

\(^9\) Patrick Andrew, in a post on the 12th of March of 2018: https://tinyurl.com/andrew-distance

\(^10\) Patrick Andrew in a post on the 12th of March of 2018: https://tinyurl.com/andrew-distance
and the updated target, we calculated for each segment: 100 – PEM%. We still call the output “PEM%” (Figure 1, right part).

### 2.1 Selection of the Sample

A client who started a collaboration (using PEMT workflow) with Datawords in 2014 agreed to be the object for this user study. This client owns an e-commerce site where motorcycling parts and accessories are sold by many brands. It also publishes a blog on the subject, thus, we consider two text types: product descriptions, and blog.

Two completed post-editing jobs were selected from 2017 archives at Datawords, one randomly and one consciously as it was the first blog article to be post-edited. Datawords had advised against using PEMT for this text type but the client insisted on doing a test. The outcome of the test was negative: the client then hired two in-house translators for the blog for an internal mission. This user study will prove that advising against PEMT for this type of text was legitimate, but we claim that alternative solutions could have been considered.

### 2.2 Cleaning the Training Data

The translation memories (TM) available at Datawords for FR-IT contained around 104,000 TUs, and for FR-ES around 120,000. After cleaning all non-relevant TUs using Heartsome TMX Editor\(^\text{11}\) (such as number only segments or duplicates), we were left with 94,000 TUs for FR-IT and 100,000 TUs for FR-ES. We also anonymized the TMX and extracted the test data\(^\text{12}\) from the training data.

Bearing in mind that NMT is less robust in low-resource situations (Blanchon and Besancier, 2017; Omniscien, 2017; Koehn and Knowles, 2017), and assuming that the training data for a neural engine should be around 10 million words\(^\text{13}\), we are aware that large parallel ad hoc corpora are required to obtain truly comparable MT outputs between NMT and PBMT. However, for specialized domains (motorcycling) and some pairs of languages (French as source) such corpora do not exist. The attempt to build parallel ad hoc corpora in such cases is, in a real scenario, confronted with many obstacles\(^\text{14}\). We are aware that many general corpora could have been used nevertheless\(^\text{15}\), however, for this first experiment we prioritized in-domain validated data (Datawords TMs only). We acknowledge interest in using big corpora in the future to compare the results.

### 2.3 Creation of a Post-Editor Pool

To find 10 participants for the experiment in both language pairs, we contacted old colleagues for the FR-ES pair and we looked for profiles in PROZ\(^\text{16}\) and Translator’s Café\(^\text{17}\). We asked

---

\(^\text{11}\) https://github.com/heartsome/tmxeditor

\(^\text{12}\) The test data is a part of the product description the participants would post-edit. The reference is the translation Datawords delivered to the customer.

\(^\text{13}\) Faherty, 2018, in an e-mail exchange.

\(^\text{14}\) We used HTTrack Website Copier, Sketch Engine and LFAligner but the results were not satisfactory, since a lot of non-relevant content (like Cookies text, etc.) would be pulled out by the software and the alignment tasks would represent months of work, which is not acceptable in most real case scenarios.

\(^\text{15}\) For instance, DGT or UN corpora, among others available at http://opus.nlpl.eu/DGT.php

\(^\text{16}\) https://www.proz.com/
them to fill in a Google Form (an example for the Spanish participants is found in Appendix A) to obtain information on their professional profile to help us create the post-editor pool. Another survey was sent as a Microsoft form (Appendix B) at the end of the experiment to complete the data with more details about their professional profiles, career and experience. In this survey we asked about training and environment (where they work, under pressure or not, etc.) and methodological questions (how they react when finding errors in source text, productivity tools used, quality assurance [QA] tools used, and MT engines tested so far).

For FR-ES, 17 people answered the form and 34 people filled it in for the FR-IT pair. All but three had 4 years or more experience as professional translators (the rest, 1 to 4 years). The most common computer-assisted translation (CAT) tool was SDL Trados Studio. Only 5 participants had 4 years of experience or more as post-editors, 8 of them had 1 to 4 years of experience in post-editing, 6 people had less than one year of experience, and 1 person had never done a PEMT project before.

2.4 Training the 3 Models

Once the phase described in section 2.2 was completed, we performed 6 trainings to compare the outputs.

Training #1 contained only the cleaned TMX without inconsistencies in target.

Training #2 was performed with the cleaned TMX without inconsistencies in target plus the termbase (TB) from Datawords.

Training #3 was based on the cleaned TMX without inconsistencies in target, plus the cleaned TB (general or non-relevant words and non-translatables removed) plus a list of non-translatables created by cleaning the TB and the Gaps file produced by KantanMT.

Training #4 contained the cleaned TMX without inconsistencies in target, plus another version of the cleaned TB (unlike #3, we not only removed general and non-translatables but also all synonyms), plus a second more comprehensive list of non-translatables, plus a file named Rejects built thanks to the TUs rejected by KantanMT.

Training #5 contained the cleaned TMX with inconsistencies both in source and target, plus the TB (as used in #4), plus the Rejects.

Training #6 contained a merged TMX (without inconsistencies in target + Rejects), plus the cleaned TB as in #4, plus the inconsistent TUs thoroughly cleaned using Xbench.\(^\text{18}\)

The evaluation of each output was done with the BLEU\(^\text{19}\) (Papineni et al., 2011) score provided by each platform (SDL Language Cloud and KantanMT):

<table>
<thead>
<tr>
<th></th>
<th>SDL Stat</th>
<th>Kantan Stat</th>
<th>Kantan NMT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FR-ES</td>
<td>FR-IT</td>
<td>FR-ES</td>
</tr>
<tr>
<td>#1</td>
<td>52</td>
<td>64</td>
<td>51</td>
</tr>
<tr>
<td>#2</td>
<td>53</td>
<td>67</td>
<td>50</td>
</tr>
<tr>
<td>#3</td>
<td>51</td>
<td>66</td>
<td>63</td>
</tr>
<tr>
<td>#4</td>
<td>53</td>
<td>66</td>
<td>68</td>
</tr>
<tr>
<td>#5</td>
<td>51</td>
<td>66</td>
<td>65</td>
</tr>
<tr>
<td>#6</td>
<td>52</td>
<td>65</td>
<td>50</td>
</tr>
</tbody>
</table>

Figure 2 BLEU evaluation

It was observed that retraining a model resulted in a better BLEU score. Unfortunately, the access to both systems had some limitations (number of trainings, number of engines, etc.)\(^\text{20}\).

---

\(^{17}\) https://www.translatorscafe.com/cafe/

\(^{18}\) https://www.xbench.net/

\(^{19}\) [Method] to compute similarity between a human supplied \textquoteleft gold standard\textquoteright reference and the MT output string based (largely) on n-gram co-occurrence. (Way, 2018)
We use training #4 for PBMT engines (despite the highest score for FR-IT in the #2, as we prioritized the comparability between the two engines), and training #5 for the Neural engine given the significantly higher score for FR-IT (and considering there was not another neural engine with which we should guarantee comparability).

Tercom was also used to calculate HTER metric\textsuperscript{21} (Snover et al., 2006). The corresponding results are presented in the next section.

3 Analysis

Thanks to the data collected through the 20 Qualitivity reports the participants delivered, we were able to proceed with multiple analysis per file, language and MT system:
- Comparison between PEM% and speed, compared to BLEU\textsuperscript{22} and HTER scores per file.
- Correlation between edit distance and speed
- Correlation between PEM% and word count
- Correlation\textsuperscript{23} between edit distance and word count
- Human evaluation (HE), through ranking\textsuperscript{24}, of the 3 MT outputs, and inter-rater agreement

<table>
<thead>
<tr>
<th>MT system: file 1</th>
<th>Average seconds</th>
<th>Average PEM%</th>
<th>BLEU score</th>
<th>HTER score</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Italian</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SDL</td>
<td>32.65</td>
<td><strong>9.85%</strong></td>
<td>69.29</td>
<td><strong>0.289</strong></td>
</tr>
<tr>
<td>KantanMT Stat</td>
<td><strong>23.62</strong></td>
<td>10.07%</td>
<td>66.82</td>
<td>0.297</td>
</tr>
<tr>
<td>KantanMT NMT</td>
<td>25.71</td>
<td>15.55%</td>
<td>58.04</td>
<td>0.396</td>
</tr>
<tr>
<td>Google Translate</td>
<td></td>
<td></td>
<td>42.20</td>
<td>0.492</td>
</tr>
<tr>
<td><strong>Spanish</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SDL</td>
<td>36.45</td>
<td><strong>9.81%</strong></td>
<td><strong>61.97</strong></td>
<td><strong>0.302</strong></td>
</tr>
<tr>
<td>KantanMT Stat</td>
<td><strong>28.36</strong></td>
<td>10.75%</td>
<td>59.63</td>
<td>0.363</td>
</tr>
<tr>
<td>KantanMT NMT</td>
<td>42.16</td>
<td>19.57%</td>
<td>41.43</td>
<td>0.488</td>
</tr>
<tr>
<td>Google Translate</td>
<td></td>
<td></td>
<td>32.59</td>
<td>0.573</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>MT system: file 2</th>
<th>Average seconds</th>
<th>Average PEM%</th>
<th>BLEU score</th>
<th>HTER score</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Italian</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SDL</td>
<td>41.00</td>
<td><strong>29.32%</strong></td>
<td>32.12</td>
<td><strong>0.649</strong></td>
</tr>
<tr>
<td>KantanMT Stat</td>
<td><strong>39.15</strong></td>
<td>30.77%</td>
<td>30.83</td>
<td>0.663</td>
</tr>
<tr>
<td>KantanMT NMT</td>
<td>39.61</td>
<td>42.53%</td>
<td>22.84</td>
<td>0.851</td>
</tr>
<tr>
<td>Google Translate</td>
<td></td>
<td></td>
<td>28.57</td>
<td>0.714</td>
</tr>
<tr>
<td><strong>Spanish</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SDL</td>
<td>54.11</td>
<td><strong>26.99%</strong></td>
<td>27.84</td>
<td><strong>0.698</strong></td>
</tr>
<tr>
<td>KantanMT Stat</td>
<td>49.10</td>
<td>31.95%</td>
<td><strong>30.22</strong></td>
<td><strong>0.698</strong></td>
</tr>
<tr>
<td>KantanMT NMT</td>
<td><strong>48.78</strong></td>
<td>46.56%</td>
<td>21.22</td>
<td>0.852</td>
</tr>
<tr>
<td>Google Translate</td>
<td></td>
<td></td>
<td>29.60</td>
<td>0.697</td>
</tr>
</tbody>
</table>

Figure 3 Comparison: PEM%, speed, BLEU and HTER scores

\textsuperscript{20} Only KantanMT allowed retraining (epochs), and only SDL Language Cloud has the *Adaptive* feature. These functionalities were not used, as they were not comparable between systems.

\textsuperscript{21} http://www.cs.umd.edu/~snover/tercom/

\textsuperscript{22} This time calculated via https://www.letsmt.eu/Bleu.aspx to allow a file per file comparison.

\textsuperscript{23} All correlations have been calculated with the Excel formula: https://tinyurl.com/correl-XLSX

\textsuperscript{24} The classification of evaluation types is based on TAUS webinar on 11\textsuperscript{th} of October of 2017.
For both languages and files, we observed a higher PEM% for the NMT model. In general, BLEU and HTER scores were consistent with PEMT%, as one of the two PBMT systems always had the highest BLEU score and either the lowest PEM%, or the highest speed. Namely, SDL showed better performance regarding automatic scores and PEM%, whereas KantanMT Stat showed an improved processing speed.

Although there are inconclusive results regarding the speed per segment on SDL outputs, it must be said, however, that the very first segment of both files was provided by this model. One explanation of the higher times required (but reduced PEM%) could be that after opening the editor in Trados Studio the participant still needed some time before starting the actual post-editing (to switch on Flashback recorder, to check Qualitivity is running, etc.)

TER is considered good enough under the threshold of 30 and, regarding the first file, the best output is produced by SDL engine.

<table>
<thead>
<tr>
<th></th>
<th>SDL</th>
<th>KantanMT Stat</th>
<th>KantanMT NMT</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>ES file 1</em></td>
<td>8.62</td>
<td>9.47</td>
<td>17.85</td>
</tr>
<tr>
<td><em>IT file 1</em></td>
<td>7.42</td>
<td>8.27</td>
<td>13.33</td>
</tr>
<tr>
<td><em>ES file 2</em></td>
<td>28.01</td>
<td>33.91</td>
<td>51.27</td>
</tr>
<tr>
<td><em>IT file 2</em></td>
<td>30.30</td>
<td>32.05</td>
<td>45.99</td>
</tr>
</tbody>
</table>

Figure 4 Edit distance score per engine and file

When we compare average edit distance to average time spent we find a remarkably high correlation for KantanMT systems:

<table>
<thead>
<tr>
<th></th>
<th>SDL</th>
<th>KantanMT Stat</th>
<th>KantanMT NMT</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Spanish file 1</em></td>
<td>40.82%</td>
<td>78.07%</td>
<td>64.23%</td>
</tr>
<tr>
<td><em>Italian file 1</em></td>
<td>58.98%</td>
<td>73.38%</td>
<td>67.39%</td>
</tr>
<tr>
<td><em>Spanish file 2</em></td>
<td>49.65%</td>
<td>76.03%</td>
<td>65.93%</td>
</tr>
<tr>
<td><em>Italian file 2</em></td>
<td>53.22%</td>
<td>76.98%</td>
<td>63.61%</td>
</tr>
</tbody>
</table>

Figure 5 Correlation: edit distance and speed

We also analyzed how PEM% related to the wordcount per segment. We found that NMT outputs did imply a higher post-editing effort in general, but we failed to find any correlation between length and PE effort, as shown in the example below:

Logically, we did obtain a closer relation between edit distance and wordcount (length), especially for file 2:

---

25 As has been confirmed thanks to Flashback recordings.
26 https://tinyurl.com/kantan-ter

71
Finally, the HE showed a preference for SDL output, which relates well with the good results shown by automatic metrics for this engine, but contrasts with speed mentioned above for KantanMT Stat. As expected, it yielded a complete refusal of MT outputs for file 2:

<table>
<thead>
<tr>
<th>Language</th>
<th>MT system</th>
<th>Preference by participants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Italian</td>
<td>SDL</td>
<td>30.43%</td>
</tr>
<tr>
<td></td>
<td>Kantan Stat</td>
<td>19.40%</td>
</tr>
<tr>
<td></td>
<td>Kantan NMT</td>
<td>14.07%</td>
</tr>
<tr>
<td></td>
<td>None</td>
<td>36.10%</td>
</tr>
<tr>
<td>Spanish</td>
<td>SDL</td>
<td>39.10%</td>
</tr>
<tr>
<td></td>
<td>Kantan Stat</td>
<td>22.47%</td>
</tr>
<tr>
<td></td>
<td>Kantan NMT</td>
<td>10.70%</td>
</tr>
<tr>
<td></td>
<td>None</td>
<td>27.73%</td>
</tr>
</tbody>
</table>

Inter Rater Reliability of the analysis was calculated in Excel for HE on a segment level. The number of coincidences were added up for each two persons and divided by the number of questions\textsuperscript{27} (75 for the product description and 55 for the blog\textsuperscript{28}):

<table>
<thead>
<tr>
<th>Language</th>
<th>File 1</th>
<th>File 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Italian</td>
<td>47.67%</td>
<td>50.95%</td>
</tr>
<tr>
<td>Spanish</td>
<td>54.01%</td>
<td>50.06%</td>
</tr>
</tbody>
</table>

\textsuperscript{27} https://www.youtube.com/watch?v=fq_LNTPgVF8
\textsuperscript{28} The reason why the number of questions does not correspond to the number of TUs as announced in 2 Method is because one TU for the product description and two for the blog article are, for the purposes of human evaluation, considered as one translation unit (though technically they are formed by 2 segments). It is shown in Appendix C.
4 Results

Independently of MT system and target language, the blog text shows, as expected, less promising results than the product description, since it is a creative text type with longer, non-repetitive sentences. Even for an in-domain setting, PBMT gives better output regarding PE effort, as shown in this new case study, where the two statistical engines (SDL and KantanMT) perform better than Google Translate (not trained with in-domain data). Indeed, it is the first time that NMT is studied for motorcycling domain with French as source language, and the fact that PBMT outperforms NMT in such conditions is due to the fact that training data was insufficient and retraining (epochs) was not performed. Nevertheless, even for such a niche domain the results with PBMT are not as positive as has been found in the past in other domains (journalism, medical, etc.). We claim that niche domains are often the most difficult cases to handle by LSPs and corporations, but have received little attention by the research community until now.

Both groups of raters show over 50% agreement on the fact that none of the 3 models perform well enough for file 2. According to the HE, SDL output is the best one for file 1 (though Italian raters are more exacting). Considering automatic evaluation, except for the Spanish SDL engine, the edit distance score is clearly over 30, where post-editing is no longer considered worth the work.

Through the form the participants filled in at the end of the task, we can see that 50% of them found errors in the source text. Thus we consider that, when deciding whether a project is apt for PEMT, not only the type of text and the domain are relevant, but also a “quality plan” engaged between customer and LSP\(^\text{29}\). This quality plan should consider a list of characteristics of the project, for instance: the final function of the translation (Skopos), and the linguistic quality of the source text (a precise analysis of amount of new vocabulary, lexical density, etc.). In this sense, the LSP could articulate a better strategy upstream when handling niche industries and/or low-resourced languages, not only by preparing general parallel corpora beforehand but also by stating the obstacles these circumstances represent in the TSP-client agreement.

It must also be said, that the post-edited outputs were at a later stage evaluated by a group of annotators (10 project managers per target language) at Datawords, who all had higher education in translation. Nonetheless, the results are being analyzed at the time of the submission of this paper and will be presented at ASLING TC40.

Finally, it is worth highlighting that one of the most visible and immediate results of this user study is the consideration by Datawords to look for a new MT provider. Even though neither of the two tested providers is the one currently contracted by Datawords, we noticed a better communication strategy from them: they informed us about how deleting synonyms from the TB could improve the quality of the MT output, among other pieces of information not only useful for the trainings themselves (which in SAS options are performed by the MT provider itself), but also relevant for Datawords when preparing the resources that will be sent to train the engines. In the past, this lack of information has contributed to unsatisfactory MT output because Datawords had not been given details on how to better prepare the resources (namely TMX and TBX).

5 Further work

As stated in this study, foreseeing and evaluating the feasibility of PEMT for a given project is a complex task. Indeed, it implies TM management, MT training (and retraining\(^\text{30}\)) and

\(^{29}\) As also suggested in ISO 18587:2017, Annex D: “Client-TSP agreements and project specifications”.

\(^{30}\) Retraining is shown to greatly improve performance when input sentences are taken from the same domain. (Denkowski, 2012).
evaluation, quality estimation, pre- and post-editing, and other tasks such as evaluation of software. These competences contribute to a professional profile that is rarely taken into consideration by academia and industry until now: the “CAT Tool Consultant”, similar to the “Paralinguist” (Van Ess-Dykema et al., 2010; Van Ess-Dykema, 2011). We shall look deeper into matters such as TM/MT integration, quality evaluation models (MQM, for instance), source file analysis, automatic post-editing (APE)… to ultimately define the competences of such a profile and include them in translator curricula.

We believe ISO 18587:2017 should be further analyzed in the context of real scenarios at some LSP and be compared with PE training in educational institutions. Therefore, a lot of work remains to be done on the post-editor profile and set of skills and competencies.

With a view to converge translation practice and theory and define a framework on how to come to a fully informed decision when selecting a PEMT strategy, we shall further the present research with similar experiments in PEMT.

Aknowledgements

Ideas and results presented in this paper are part of Clara Ginovart Cid’s PhD research, conducted at Pompeu Fabra University, under the supervision of Pr. Carme Colominas and Pr. Antoni Oliver (Universitat Oberta de Catalunya), in collaboration with Datawords Datasia, under the supervision of Marina Frattino, supported through the Industrial Doctorate Programme. We would like to show our gratitude to the 20 post-editors and the 20 annotators who made this case study possible.

References


Faherty, Louise. 2018. KantanMT PhD study UPF. [Personal e-mail].


31 EAGLES (Starlander, 2015)
33 To determine the level of repetitions, new vocabulary, errors, and therefore include the source text NTIs (negative translatability indicators), as proposed in O’Brien (2006), among the criteria to decide if a project is suitable for PEMT or not.

Harris, Kim, Alan Kenneth Melby, Attila Görög, & Serge Gladkov. 2015. Multidimensional Quality Metrics. German Research Center for Artificial Intelligence (DFKI) and QTLanchnPad. Berlin, Germany.


ISO 17100. 2015. Translation services — Requirements for translation services. Switzerland.

KantanMT. 2015. What is Translation Error Rate (TER)? Retrieved: https://tinyurl.com/kantan-ter


Appendix A – Questionnaire to select post-editors (Google form)

This questionnaire should help select a limited number of participants for a case study in the frame of an Industrial Doctorate. PhD student is Clara Ginovart and the thesis will be developed in collaboration between Datawords Datasia and Pompeu Fabra University.

Email address*
Have you studies in translation?*
  Yes
  No
  Other:
Are you a native Spanish speaker?*
  Yes
  No
Do you have the language pair French-Spanish?*
  Yes
  No
How many years have you been working as a professional translator?*
  Less than 1
  1 to 3
  More than 3 years
In what 3 fields do you have most experience?*
  Financial translation
  Food
  Tourism
  Marketing
  Psychology or religion
  Robotics
  Automotive
  Natural sciences
  Sports
  Patents
  Localization
  Medical translation
  Technical translation
  Literary translation
  Legal translation
  Audiovisual
  Other:
Do you have experience in post-editing machine translation? *
  Yes, less than one year
  No
  Yes, 1 to 2 years
  Yes, 2 years or more

Appendix B – Retrospective survey (Microsoft form)

Survey for participants in Post-Editing exercise MB DW UPF PEMT Task 2018

1. Full name
   Please enter your full name:

2. E-mail
   Please enter your e-mail:

3. Studies
   If you have followed some other courses, please specify them in "Other" field.
   - University Degree in Translation
   - University Degree in another domain
   - Certificate in Post-Editing
4. What are your language pairs?
You can use symbol ">" to indicate translation direction, such as "French>Italian":

5. Years of experience as professional translator
- < 1
- 1 - 4
- 4 or more

6. Status as a translator
- Freelance
- In-house in a corporate or firm
- In-house in an LSP
- Student
- Other:

7. Does your institution offer post-editing training to you?
The university where you study, the firm where you work at, the LSP you work for, etc.
- Always
- Often
- Sometimes
- Hardly ever
- Never

8. Where do you most often work?
- At home
- In a private office
- In a co-working or other shared office
- In a library
- Other:

9. Do you often work under time pressure?
You can give details on the requested output per hour/day/week.

10. Do you use any tool or strategy measure your productivity?
Such as keeping track of the words translated per day or per hour. If your answer is "Yes", please explain which tool or strategy you use.

11. How often do you find errors in source text?
- Always
- Often
- Sometimes
- Hardly ever
- Never

12. When you find an error in source text, do you flag it to your customer?
You can describe an example in "Other" field.
- Yes
- No
- It depends
- Other:
13. CAT tool most used or you are most familiar with:

- Wordfast
- SDL Trados Studio
- OmegaT
- None
- Memsource
- Wordbee
- MemoQ
- Other:

Appendix C – Translation Units formed by 2 segments

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>297</td>
<td>Poids :</td>
</tr>
<tr>
<td>298</td>
<td>1450g +/-50g</td>
</tr>
<tr>
<td>299</td>
<td>Poids :</td>
</tr>
<tr>
<td>300</td>
<td>1450g +/-50g</td>
</tr>
<tr>
<td>301</td>
<td>Poids :</td>
</tr>
<tr>
<td>302</td>
<td>1450g +/-50g</td>
</tr>
<tr>
<td>303</td>
<td>Merci d'indiquer quelle traduction vous préférez avec « 1 », « 2 » ou « 3 », si aucune, rentrez « 0 ».</td>
</tr>
</tbody>
</table>

Figure 11 Divided translation unit 1 - file 1

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>125</td>
<td>Et bien sûr, sous la pluie, on baigne rapidos dans son jus.</td>
</tr>
<tr>
<td>126</td>
<td>Normal.</td>
</tr>
<tr>
<td>127</td>
<td>Et bien sûr, sous la pluie, on baigne rapidos dans son jus.</td>
</tr>
<tr>
<td>128</td>
<td>Normal.</td>
</tr>
<tr>
<td>129</td>
<td>Et bien sûr, sous la pluie, on baigne rapidos dans son jus.</td>
</tr>
<tr>
<td>130</td>
<td>Normal.</td>
</tr>
<tr>
<td>131</td>
<td>Merci d'indiquer quelle traduction vous préférez avec « 1 », « 2 » ou « 3 », si aucune, rentrez « 0 ».</td>
</tr>
</tbody>
</table>

Figure 12 Divided translation unit 2 - file 2

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>204</td>
<td>Mon avis :</td>
</tr>
<tr>
<td>205</td>
<td>Du frais pour vos guibolles</td>
</tr>
<tr>
<td>206</td>
<td>Mon avis :</td>
</tr>
<tr>
<td>207</td>
<td>Du frais pour vos guibolles</td>
</tr>
<tr>
<td>208</td>
<td>Mon avis :</td>
</tr>
<tr>
<td>209</td>
<td>Du frais pour vos guibolles</td>
</tr>
<tr>
<td>210</td>
<td>Merci d'indiquer quelle traduction vous préférez avec « 1 », « 2 » ou « 3 », si aucune, rentrez « 0 ».</td>
</tr>
</tbody>
</table>

Figure 13 Divided translation unit 3 - file 2