# Machine translation evaluation made fuzzier: A study on post-editing productivity and evaluation metrics in commercial settings

**Carla Parra Escartín** carla.parra@hermestrans.com
**Manuel Arcedillo** manuel.arcedillo@hermestrans.com
Hermes Traducciones
C/ Cólquide 6, portal 2, 3. I
28230 Las Rozas, Madrid, Spain

**Abstract**

In this paper, we report on an experiment carried out in the context of a translation company. Ten translators, with diverse degrees of experience in translation and machine translation post-editing (MTPE), were assigned the same task, involving translation from scratch, fuzzy-match post-editing, and MTPE. We evaluate the MT output using traditional evaluation metrics such as BLEU and TER, correlate these measures with productivity values and study whether a fuzzy score stands up against them. Our main goal was to evaluate whether fuzzy scores can be used for evaluating MTPE, thus incorporating its familiarity and TM matching analogies to an MTPE workflow. The results of our experiment seem to support this hypothesis.

## 1 Introduction

In the last years, machine translation post-editing (MTPE) tasks have become ever more common in the translation industry. Translators and translation companies of all sizes use machine translation (MT) to increase their productivity and reduce production costs. However, it is still unclear how to assess MT output in order to verify that those goals have been met and, if applicable, determine a fair compensation for the post-editor.

Traditional automatic evaluation metrics such as BLEU (Papineni et al., 2001) and TER (Snover et al., 2006) have shortcomings such as unproven correlation with productivity gains (Callison-Burch et al., 2006), technical difficulties for their estimation by general users and lack of intuitiveness. Meanwhile, the translation industry has a well-established way of evaluating text similarity and establishing discount rates: translation memory (TM) fuzzy match scores (FMS). Based on text similarity with segments stored in the TM, each sentence to be translated receives a fuzzy score, ranging from 0% (no similar sentence was found in the TM) to 100% (an exact match was found). In the translation industry, TM matches above 75% are normally assigned a rate discount, while those segments below this threshold are paid the full rate following the general assumption that they do not yield any productivity increase.[1] Since translators and other parties involved in a translation project are familiar with this system and accept it as a valid business model, we designed an experiment to verify whether target-side FMS stand up

[1] This is an industry general practice. Plitt and Masselot (2010) also report that in a typical localization scenario TM matches below 75% are considered no-matches. However, lack of research and the peculiarities of fuzzy score calculation in each tool may allow for variations of this model.

against traditional methods of MTPE evaluation.

In the experiment, ten professional translators with diverse experience in the industry were assigned the same file to translate, involving translation from scratch, TM match post-editing, and MTPE. We recorded the time spent by each translator in each segment in order to obtain productivity values for all TM match bands, including no match segments and exact matches. We then computed several automated metrics (namely, BLEU and TER) and correlated them with productivity values. Finally, we compared the performance of these automated metrics with the performance of the target-side FMS.

The remainder of this paper is structured as follows: Section 2 summarizes traditional MT evaluation metrics (c.f. 2.1 and 2.2) and presents the metric we use in our experiments, explaining why we deem it appropriate for MT evaluation (c.f. 2.3). The details of the experiment itself are explained in Section 3. In Section 4 we present the results and analyze them. Finally, in Section 5 we summarize the findings and discuss possible future research.

## 2 Background

A common practice when negotiating MTPE discounts is to either annotate a sample of the MT output according to any of the methods described below or similar ones, or post-edit it in order to generate a reference for automatic evaluation, while possibly gathering other data such as time spent or key strokes. Since rates are normally set before handing off a project, a good MT evaluation at this stage is critical to avoid underpayments and the mistrust this situation may cause.

MT evaluation is a research field in itself. As stated by King et al. (2003), Yorick Wilks has been credited with the famous remark that "more has been written about MT evaluation than about MT itself". While it is not the purpose of this paper to revise all the proposed MT evaluation metrics, we deem it necessary to acknowledge the most commonly used metrics and assess their usability from the translation industry point of view. Subsection 2.1 summarizes human evaluation and Subsection 2.2 focuses in automatic evaluation metrics. Subsection 2.3 explains the shortcomings of traditional evaluation metrics and presents the metric we additionally used in our experiments: the target-side FMS.

### 2.1 Human evaluation

As explained by Koehn (2010), two main strategies are used in evaluation campaigns: fluency and adequacy, and ranking of translations. When human judges are asked to assess the fluency and adequacy of MT output, the task consists of assigning a score from one to five on these two criteria. Fluency assesses whether the text is fluent in the target language. It refers to grammaticality, correctness and idiomatic word choices. Adequacy, on the other hand, assesses whether the output sentence conveys the same meaning as the input sentence.

As pointed out by Koehn (2010), "these definitions are very vague, and it is difficult for evaluators to be consistent in their application". Callison-Burch et al. (2007) also point this out: "No instructions are given to evaluators in terms of how to quantify meaning, or how many grammatical errors (or what sort) separates the different levels of fluency. Because of this many judges either develop their own rules of thumb, or use the scales as relative rather than absolute." To overcome these issues, the "translation ranking" approach may be taken. In this kind of evaluation, human judges are given the output of different systems and are asked to choose the best one. This approach is based on the judges' perception of usefulness in terms of savings and productivity, rather than on usefulness itself. In a translation and productivity test carried out by Autodesk in 2011,[2] a clear mismatch between perception of productivity and actual productivity was found.

[2] http://langtech.autodesk.com/productivity.html

An alternative evaluation methodology is the one proposed by Hurtado Albir (1995). This methodology is widely used in translator training courses. It distinguishes different types of errors such as orthotypographical, grammatical, or semantic errors. More recent proposals are the TAUS Dynamic Quality Framework (DQF),[3] the Multidimensional Quality Metrics (MQM) Framework proposed by the QTLaunchPad project (Burchardt and Lommel, 2014), and their harmonized version, recently released (Lommel, 2015). However, since human evaluation (as opposed to automatic evaluation) is costly[4] and time consuming, these methods are not always feasible.

### 2.2 Automatic evaluation

Throughout the years, several automatic evaluation metrics have been proposed. These can be grouped in lexical, syntactic and semantic evaluation metrics depending on the linguistic level at which they operate.

Lexical evaluation metrics assess the lexical similarity between the MT output and one or several references. These metrics assume that the usefulness of MT is related to its proximity or similarity to the reference(s). This assumption is most informative when such reference(s) are post-edits of the MT output, rather than previous translations.

They can be further classified into metrics measuring the edit distance, the lexical precision, the lexical recall and the F-Measure. Edit distance is measured by WER (Word Error Rate) (Nießen et al., 2000) and TER (Translation Edit Rate) (Snover et al., 2006, 2009). Lexical precision is measured by BLEU (Bilingual Evaluation Understudy) (Papineni et al., 2001) and NIST (Doddington, 2002). ROUGE (Lin and Och, 2004) assesses the lexical recall and PER (Position-independent Word Error Rate) (Tillmann et al., 1997) assesses the lexical precision and recall. Finally, $GTM_e$ (Melamed et al., 2003) and METEOR (Lavie and Agarwal, 2005; Denkowski and Lavie, 2010) are based on the F-Measure.

MT evaluation tools such as Asiya (Giménez and Màrquez, 2010) additionally offer syntactic and semantic similarity evaluation metrics. The syntactic metrics capture similarities over shallow-syntactic structures, dependency relations and constituent parse trees. The semantic metrics intend to capture similarities over named entities, semantic roles and discourse representations. Both the syntactic and the semantic metrics require the usage of additional Language Resources and Tools (LRT), such as parsers, corpora, and specific packages.

All these metrics share the advantage of being fast and cheap to obtain. However, they still remain obscure for laymen and professional translators not familiarized with MT evaluation research, and typically are not designed with MTPE in mind.

### 2.3 Fuzzy match scores for MT evaluation

Although BLEU is a well-established evaluation metric and it is extensibly used in MT engine development cycles and the evaluation of raw MT output as a final product, it has also received criticism (Callison-Burch et al., 2006; Koehn, 2010; Bechara, 2013). A general rule of thumb in MT evaluation is that BLEU scores above 30 reflect understandable translations, while scores over 50 can be considered good and fluent translations (Lavie, 2010). However, the usefulness of "understandable" translations is questionable in the case of MTPE, since post-editors do not depend on MT to understand the meaning of the source text. Also, how should one interpret a BLEU score improvement from 45 to 50 in terms of productivity? Taking another widely-used metric, does a TER value of 40 justify any discount? The vast majority of translators (and even MT researchers) would probably be unable to answer these questions. And yet, they would probably instantly acknowledge that TM fuzzy matches of 60% are not worth editing, while

[3] `https://evaluate.taus.net/evaluate/about`

[4] See, for instance, Section 3, "Costs", in Burchardt and Lommel (2014).

they would probably consider it fair to accept discounts for 80% matches. We believe MTPE tasks would benefit greatly from the familiarity of source-side FMS applied to MTPE evaluation as a target-side FMS.[5]

Organizations such as TAUS have already proposed alternative models of MT evaluation which use FMS, such as the "MT Reversed Analysis".[6] Also, Zhechev (2012) created a metric combining character and word-based FMS and reported it as having the greatest correlation with productivity gain. In this experiment, bearing in mind that each CAT tool uses a different (and generally unknown) algorithm for their source-side FMS, we use the fuzzy value computed by Okapi Rainbow[7] in its Translation Comparison feature. We use this tool because it is freely available and it natively supports the most common bilingual formats, hence potentially improving transparency and usability for all parties involved in a translation project. This FMS is based on the Sørensen-Dice coefficient (Sørensen, 1948; Dice, 1945) using 3-grams. Equation 1 illustrates how the FMS is computed by Okapi Rainbow. An important difference with other automatic metrics is that it does not depend on traditional tokenization: instead of word n-grams, it computes character n-grams. Segments are split into a list of characters, which is then used to compute 3-grams and string size.[8]

$$\frac{2 * (MT\_output \cap PE\_output)}{size\_MT\_output + size\_PE\_output)} * 100 \quad (1)$$

The MT output (or TM output in the case of TM matches) was compared against the post-edited output of each translator. This means that each segment has ten separate scores, one for each translator. It is also worth noting here that translators faced either MT output or TM output –or no suggestion at all in segments selected for translation from scratch–, so the original output is unambiguous.

## 3 Experiment settings

The experiment was based on a pilot experiment (Parra Escartín and Arcedillo, 2015) which seemed to suggest that fuzzy scores might be as good as or even better than BLEU and TER scores for evaluating MTPE. Similarly to what Federico et al. (2012) did, we replicated a real production environment.

Ten in-house translators were asked to translate the same file from English into Spanish. We asked them to translate the text using one of their most common CAT tools, memoQ.[9] This tool was chosen because it allows keeping track of the time spent in each segment. We disregarded using other tools which also record time and other useful segment-level indicators, such as keystrokes in PET (Aziz et al., 2012) and iOmegaT (Moran et al., 2014), or MTPE effort in MateCat (Federico et al., 2012) because they are not part of the everyday tools of the translators involved. Translators were only allowed to use the TM, the terminology database and the MT output included in the translation package. We disabled all other memoQ's productivity enhancing features such as predictive text, sub-segment leverage and automatic fixing of TM matches to allow for better comparisons with translation environments which may not offer similar features.

[5] Hereinafter, FMS will refer exclusively to target-side FMS unless otherwise specified.

[6] https://www.taus.net/think-tank/best-practices/postedit-best-practices/pricing-machine-translation-post-editing-guidelines

[7] http://okapi.opentag.com/

[8] For further information, see the open source code available at: https://bitbucket.org/okapiframework/.

[9] The version used was memoQ 2015 build 3.

### 3.1 Test set selection

The selection criteria for the text to be translated were the following:

1. Belong to a real translation request.
2. Originate from a client for which our company owned a customized MT engine.
3. Have a word volume capable of engaging translators for several hours.
4. Include significant word counts for each TM match band (i.e., exact matches, fuzzy matches and no-match segments.

We used the word count analysis already available in our company's archive to find a translation project which met our criteria. In this case, the existing word count analysis came from SDL Trados Studio.[10] The original source text had over 8,000 words and was part of a software user guide. To avoid skewing due to the inferior typing and cognitive effort required to translate the second of two similar segments, repetitions and internal leverage segments were filtered out.

When transferring the project to memoQ (the CAT tool chosen for the experiment), we encountered a quite different word count. We expected some discrepancies in this new analysis, but they turned out to be bigger than we preliminary thought. Table 1 shows the word count distribution reported by both tools. As can be observed in Table 1, memoQ's word count for the 95–99% fuzzy band ended up being significantly smaller despite our efforts to select a text with a more balanced distribution. This is evidence of a well-known problem in the translation industry: the different algorithms used by each tool for computing word counts and calculating fuzzy scores.

| | SDL Trados Studio | | memoQ | |
|---|---|---|---|---|
| **TM match** | **Words** | **Segments** | **Words** | **Segments** |
| 100% | 1243 | 95 | 1226 | 94 |
| 95–99% | 1044 | 55 | 231 | 21 |
| 85–94% | 747 | 43 | 1062 | 48 |
| 75–84% | 608 | 42 | 696 | 42 |
| No Match | 3388 | 233 | 3804 | 263 |
| **Total** | **7030** | **468** | **7019** | **468** |

Table 1: Word count as computed by SDL Trados Studio and memoQ.

The differences in word count unfortunately created a less informative 95–99% fuzzy band. On the other hand, the increased number of no-match words (from 3388 to 3804) provided us with a more solid sample for comparing MTPE and translation throughputs. In order to make this comparison, we randomly divided the no-match band in two halves using the test set generator included in the *m4loc* package.[11]. One half was translated from scratch, and the other half was machine translated and subsequently post-edited.

### 3.2 Machine translation system used

To generate the raw MT output we used a customized Systran's RBMT engine.[12] This system was selected because it is the one we normally use for MTPE for this client. It can be

[10] The version used was SDL Trados Studio 2014 SP1.

[11] *m4loc* is an open source project initiative to "combine investments to efficiently leverage the potential that Moses promises for localization" (Ruopp, 2010) For more information and access to the code see: `https://code.google.com/p/m4loc`

[12] Systran 7 Premium Translator was used.

considered a mature engine since at the time of the experiment it had been subject to ongoing in-house customization for over three years and boasted a consistent record for productivity enhancement. The customization includes dictionary entries, software settings, and pre- and post-editing scripts. Although Systran includes a statistical machine translation (SMT) component, this was not used in our experiment. This decision was taken based on our past experience. In our experiments using this component, it seemed to negate the strengths of RBMT engines (such as predictability and orthotipographic consistency), while not incorporating enough improvements to compensate for them. While we are aware of recent advances in this topic, this component is not part of our usual MTPE preparation workflow and thus we chose not to use it in our experiment.

### 3.3 Translators participating in the experiment

Ten translators were engaged in the experiment. Two carried out the task as part of a pilot experiment (Parra Escartín and Arcedillo, 2015) and eight more were engaged with the aim of verifying the initial findings. A preliminary survey was sent to all translators to gather information on their experience. Translators were asked to provide their years of experience in translation and MTPE, their experience translating and/or MTPE texts from this client, their opinion on MT (positive or negative), and their opinion on CAT tools (positive or negative). Table 2 summarizes the results of our survey. Translators are sorted in descending order according to their combined experience in translation and MTPE.

| | **Trans. exp.** | **MTPE exp.** | **Client exp.** | **MT opinion** | **CAT opinion** |
|---|---|---|---|---|---|
| **TR1** | 5 | 3 | Yes | Positive | Positive |
| **TR2** | 5 | 3 | Yes | Positive | Positive |
| **TR3** | 5.5 | 2 | Yes | Positive | Positive |
| **TR4** | 5 | 1 | Yes | **Negative** | Positive |
| **TR5** | 5 | 0 | No | Positive | Positive |
| **TR6** | 5 | 0 | No | Positive | Positive |
| **TR7** | 2 | 1 | Yes | Positive | Positive |
| **TR8** | 1.5 | 1.5 | Yes | Positive | Positive |
| **TR9** | 2 | 0 | No | **Negative** | Positive |
| **TR10** | 1 | 0 | No | Positive | Positive |

Table 2: Overview of translator's experience (measured in years) and opinion on MT and CAT.

As some translators may be biased against MT, we deemed it necessary to gather their view on it as positive or negative, to see if this may also have an impact in the results. Only two translators expressed that they did not like working with MT or MTPE, even though they acknowledged that high quality MT output generally enhances their productivity.

## 4 Results and discussion

The ten packages received from the translators were analyzed individually in order to extract the time spent in each segment and calculate automated evaluation metrics. Segments with exceptionally long times (above 10 minutes) and those with unnaturally high productivity (above 5 words per second) were considered outliers and were discarded for further analysis. The few exceptionally long times registered can be explained by the translator not "closing" the segment during long pauses or lunch breaks. Since the two segments with highest recorded times (above one hour) belong to the translators who openly hold a grudge against MT, it is tempting to interpret this as an attempt to pervert the results. However, since the rest of their samples showed no other data anomalies and the time at which those exceptions were edited matches

translators' normal lunch breaks, it can be safely assumed that those two cases were exceptions due to carelessness rather than deliberate data tampering.

A possible explanation for the segments with unnaturally high productivity might be that the translator actually proofread the translation when having the cursor placed in the previous segment. Some of these segments were 100% matches and thus it seems logical that they only opened the segment to confirm it and move to the next one. However, we lack a way to confirm this hypothesis.

Data series of Translators 5 and 8 also showed questionable results. Translator 5's MT post-edited output quality showed clear signs of under-editing, which would not have reached the minimum quality standards required by the client in a real project. This under-editing can also be seen when comparing MTPE sample's automated metrics across translators. As table 3 shows, values for Translator 5 differ significantly from the rest. Closer analysis of the fuzzy and translation-from-scratch final text also showed significantly more errors than in the rest of translators. Although it may prove an interesting starting point for further research on MTPE and translation errors, we have omitted the analysis of this data series for the purpose of this paper.

| | **TR-1** | **TR-2** | **TR-3** | **TR-4** | **TR-5** | **TR-6** | **TR-7** | **TR-8** | **TR-9** | **TR-10** |
|---|---|---|---|---|---|---|---|---|---|---|
| **BLEU** | 64.24 | 66.37 | 64.23 | 63.40 | **78.16** | 69.57 | 66.42 | 63.56 | 65.88 | 68.43 |
| **TER** | 22.14 | 20.98 | 22.32 | 24.52 | **12.82** | 18.89 | 19.74 | 21.40 | 19.77 | 18.84 |
| **FMS** | 87.29 | 87.74 | 87.05 | 86.30 | **93.06** | 88.53 | 88.40 | 87.62 | 88.85 | 89.09 |

Table 3: Automated metrics for the MTPE sample per translator.

Translator 8's anomaly was detected when studying correlation values between automated metrics and productivity. In this case, it was significantly lower than for the rest of translators (see Table 9 in Section 4.3). Upon further enquiries, Translator 8 admitted to have enabled the predictive text feature which had been originally disabled from the hand-off package. Again, although extremely interesting for future research, for the purpose of this paper Translator 8's data series was discarded.

One last remark about possible anomalies is the generally low productivity of the 95–99% TM match band when compared to contiguous TM match bands (Table 4). In the vast majority of these segments there were no textual differences between the match stored in the TM and the content to be translated. This is illustrated in Example 2.

(2) **Source in TM**: If protection has been disabled manually, you can enable it in the application settings window.
**Target in TM**: Si se ha desactivado la protección manualmente, puede activarla en la ventana de configuración de la aplicación.
**New Source**: If protection has been disabled manually, you can <**1**>enable it in the application settings window <**/1**>.
**New Target**: Si se ha desactivado la protección manualmente, puede <**1**>activarla en la ventana de configuración de la aplicación<**/1**>.

The results obtained show that the time needed to perform these inline tag edits does not match what would be expected of an equivalent 95% TM match due to text differences. Even though it is possible to adjust the fuzzy score assigned by CAT tools to these segments via formatting penalties (as in SDL Trados Studio) or by using tag weights (as in memoQ), the impact of inline tag editing in translation throughput has not yet been extensively researched. For this reason, we have omitted the 95–99% band in all following analysis.

### 4.1 Productivity report

Although a standard reference of 313–375 words per hour (2500–3000 words per day) is usually taken as a baseline for productivity comparisons, we calculated each translator's translation-from-scratch throughput individually. As expected, it varied greatly, ranging from 473 to 1701 words per hour (Table 4). It is remarkable that Translator 1 was even able to translate from scratch faster than when post-editing 75–84% TM matches. In a real project, a rate discount would have been applied to those matches, which apparently would have been unfair for this translator.

Not depending on standard reference values also avoided misleading interpretations of MTPE throughputs. For example, comparing the average 1329 words per hour with a standard reference of 375 words per hour would show productivity gains of around 250%. However, based on the translation-from-scratch throughputs obtained, none of our translators reached such numbers, as can be seen in Table 5, which shows a more realistic average gain of "just" 24%. This data confirms the importance of having a translation reference for each sample instead of relying on standard values (Federico et al., 2012).

Table 4 reports the productivity (words per hour) obtained for each translator in each band. The words per hour average is provided in the last column.

| | **Seg.** | **Words** | **TR-1** | **TR-2** | **TR-3** | **TR-4** | **TR-6** | **TR-7** | **TR-9** | **TR-10** | **Avg.** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **100%** | 94 | 1226 | 3277 | 2942 | 1894 | 1767 | 1579 | **4039** | 2798 | 1395 | **2461** |
| **95–99%** | 21 | 231 | **2642** | 2625 | 1476 | 1299 | 963 | 2011 | 1133 | 1543 | **1477** |
| **85–94%** | 48 | 1062 | **2960** | 2248 | 1660 | 1678 | 1232 | 2164 | 2429 | 1012 | **1923** |
| **75–84%** | 42 | 696 | **1592** | **1592** | 1372 | 1140 | 1019 | 1342 | 1576 | 741 | **1297** |
| **MTPE** | 131 | 1890 | **1804** | 1743 | 1369 | 1141 | 922 | 1481 | 1433 | 739 | **1329** |
| **Trans.** | 132 | 1914 | **1701** | 1319 | 993 | 916 | 933 | 1236 | 1223 | 473 | **1099** |

Table 4: Productivity achieved per translator and match band in words per hour.

We additionally computed the productivity gain achieved by each translator in each band and the average gain across translators. Equation 3 shows the formula used for calculating productivity gains, where $PE\_Throughput$ is the productivity achieved by one translator in words per hour when post-editing MT output or TM matches, and $TRA\_Throughput$ is the productivity achieved by that same translator when translating from scratch (taken from productivity values in Table 4). Table 5 reports the productivity gains obtained.

$$(\frac{PE\_Throughput}{TRA\_Throughput} - 1) * 100 \tag{3}$$

| | **TR-1** | **TR-2** | **TR-3** | **TR-4** | **TR-6** | **TR-7** | **TR-9** | **TR-10** | **Avg.** |
|---|---|---|---|---|---|---|---|---|---|
| **MTPE gains** | 6.06 | 32.15 | 37.78 | 24.62 | -1.23 | 19.80 | 17.20 | 56.34 | **24.09** |

Table 5: Productivity gain percentage due to MTPE for each translator.

All translators except one (Translator 6, who experienced 1.2% productivity loss when facing MTPE) were faster in MTPE than translating from scratch, although the productivity gain varied greatly among them (from 6.06% to 56.34%). A possible explanation for the slight productivity loss experienced by Translator 6 might be that this translator had experienced an extensive period of inactivity and had barely used memoQ. The biggest productivity gain was achieved by the least experienced translator (Translator 10), while the smallest productivity gain corresponds to the most experienced translator (Translator 1), although this trend cannot be confirmed by the rest of results. Furthermore, all translators who had MTPE throughput

higher than the 75–84% band (Translators 1, 2 and 7) had previous experience in MTPE. On the contrary, both translators (6 and 9) who worked faster with 75–84% TM matches than with MT segments had no MTPE experience. This seems to point out that mature MT engines allow experienced post-editors to post-edit MT faster than the lowest TM matches.

Also, it is worth noting that the productivity in Table 4 reflects only the time spent by the translator inside the editor. Other tasks such as reading of instructions, file management, escalation of queries, self-review, quality assurance and communication with project managers and/or other parties involved in the project are not directly reflected here, but are all part of a regular project. Care should also be taken when transferring productivity gains into rate discounts, as translation rates often include tasks which are not affected by the introduction of MTPE (such as project management, file engineering or review by a second linguist). In any case, it is interesting to see that MTPE can improve productivity even when the fastest translators work with content which allow for higher-than-usual translation throughputs.

### 4.2 MT evaluation metrics

We calculated document-level values for BLEU, TER and FMS[13] (Tables 6, 7 and 8 respectively) taking the segments belonging to each band as separate documents. To allow for a more direct comparison, next to each band average we attach its corresponding average productivity gain, calculated as the average of the eight values (one for each translator) obtained according to Equation 3 above.

| | **TR-1** | **TR-2** | **TR-3** | **TR-4** | **TR-6** | **TR-7** | **TR-9** | **TR-10** | **Avg.** | **Avg. gain** |
|---|---|---|---|---|---|---|---|---|---|---|
| **100%** | 93.17 | 92.38 | 85.57 | 92.33 | 91.96 | 96.28 | 94.25 | 95.51 | 92.68 | 127.38% |
| **95–99%** | 86.51 | 89.35 | 78.96 | 81.85 | 87.91 | 85.83 | 80.32 | 92.10 | 85.35 | 66.20% |
| **85–94%** | 82.22 | 81.00 | 76.15 | 80.70 | 80.19 | 85.62 | 84.39 | 88.23 | 82.31 | 76.82% |
| **75–84%** | 70.42 | 71.98 | 66.57 | 70.34 | 73.94 | 71.50 | 70.52 | 70.36 | 70.70 | 22.52% |
| **MTPE** | 64.24 | 66.37 | 64.23 | 63.40 | 69.57 | 66.42 | 65.88 | 68.43 | 66.07 | 24.09% |

Table 6: BLEU scores for each TM match band and average productivity gain (%).

| | **TR-1** | **TR-2** | **TR-3** | **TR-4** | **TR-6** | **TR-7** | **TR-9** | **TR-10** | **Avg.** | **Avg. gain** |
|---|---|---|---|---|---|---|---|---|---|---|
| **100%** | 4.38 | 3.99 | 8.43 | 4.35 | 4.64 | 2.00 | 3.10 | 1.94 | 4.10 | 127.38% |
| **95–99%** | 8.24 | 6.37 | 14.02 | 11.50 | 6.91 | 8.31 | 12.83 | 5.30 | 9.19 | 66.20% |
| **85–94%** | 11.58 | 12.69 | 16.89 | 12.96 | 13.02 | 10.45 | 10.04 | 8.33 | 11.99 | 76.82% |
| **75–84%** | 21.65 | 21.03 | 23.74 | 21.18 | 19.08 | 20.07 | 20.86 | 20.20 | 20.98 | 22.52% |
| **MTPE** | 22.14 | 20.98 | 22.32 | 24.52 | 18.89 | 19.74 | 19.77 | 18.84 | 20.90 | 24.09% |

Table 7: TER scores for each TM match band and average productivity gain (%).

The three metrics agree that the four less experienced translators (6–10) performed less edits to the MT output and the TM matches than the four more experienced translators (1–5). The issue reported in the beginning of Section 4 about the 95–99% band not achieving the expected productivity can also be seen in these metric tables. It would be logical to expect the productivity of this band being somewhere in between the 100% and the 85–94% bands, and the three metrics indeed agree with this intuition, but in terms of productivity it was not the case.

[13] BLEU and TER were computed using Asiya (Giménez and Màrquez, 2010) and the FMS was computed using the Okapi framework. For each translator and in each segment, the reference was only the post-edited MT or TM output delivered by that same translator. In no case multiple references were used.

| | TR-1 | TR-2 | TR-3 | TR-4 | TR-6 | TR-7 | TR-9 | TR-10 | Avg. | Avg. gain |
|---|---|---|---|---|---|---|---|---|---|---|
| **100%** | 96.91 | 97.91 | 94.81 | 97.34 | 97.50 | 98.67 | 98.01 | 98.70 | 97.48 | 127.38% |
| **95–99%** | 91.62 | 92.76 | 90.71 | 91.86 | 92.43 | 92.10 | 90.00 | 95.48 | 92.12 | 66.20% |
| **85–94%** | 92.19 | 91.19 | 88.42 | 90.44 | 90.92 | 92.19 | 92.21 | 93.50 | 91.38 | 76.82% |
| **75–84%** | 85.07 | 85.21 | 83.02 | 85.62 | 87.93 | 85.6 | 85.48 | 86.83 | 85.60 | 22.52% |
| **MTPE** | 87.29 | 87.74 | 87.05 | 86.30 | 88.53 | 88.4 | 88.85 | 89.09 | 87.91 | 24.09% |

Table 8: Fuzzy scores for each TM match band and average productivity gain (%).

As explained earlier, this can be explained by the fact that the vast majority of edits required in the 95–99% band involved dealing with inline tags. Although the impact of these operations has not been researched enough, these results show that they can have a big impact in terms of productivity, slowing down the translator more than would be expected. Moreover, when calculating automated metrics, these inline tags are generally deleted to avoid their division in different tokens, thus potentially skewing the results. Instead of deleting them, when calculating the automated metrics reported in this paper we converted each tag into a unique token. This operation was clearly not enough to compensate all the effort put into tag handling, as hinted by the productivity values. More research is needed in this area to find out the appropriate weight or penalty that automated metrics should assign to inline tags. As stated earlier, we opted to treat the 95–99% band as an outlier and ignore it from further analysis.

### 4.3 Productivity vs. MT evaluation metrics correlation

To find out if the proposed FMS stands up against traditional methods of MT evaluation, we correlated the metrics in Tables 6, 7 and 8 with the productivity values in Table 4. We also added two alternatives to the FMS calculated by Okapi Rainbow: the FMS provided by SDL Trados Studio and memoQ, two of the most popular CAT tools worldwide. As can be observed in Table 9, the results indicate that BLEU, TER and the three FMS have a strong correlation with productivity. Therefore, all of them seem suitable for evaluating MTPE or TM match productivity. The anomaly created by Translator 8 using the predictive text feature is also reflected in Table 9.

| | TR-1 | TR-2 | TR-3 | TR-4 | TR-6 | TR-7 | TR-8 | TR-9 | TR-10 | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|
| $r_{\mathrm{BLEU}}$ | 0.934 | 0.952 | 0.996 | 0.934 | 0.999 | 0.925 | (0.816) | 0.995 | 0.953 | **0.961** |
| $r_{\mathrm{TER}}$ | -0.973 | -0.995 | -0.987 | -0.945 | -0.989 | -0.968 | (-0.725) | -0.984 | -0.973 | **-0.977** |
| $r_{\mathrm{fuzzy}}$ | 0.973 | 0.995 | 0.918 | 0.902 | 0.971 | 0.974 | (0.548) | 0.911 | 0.979 | **0.953** |
| $r_{\mathrm{Studio}}$ | 0.967 | 0.987 | 0.955 | 0.930 | 0.942 | 0.968 | (0.749) | 0.987 | 0.969 | **0.963** |
| $r_{\mathrm{memoQ}}$ | 0.929 | 0.962 | 0.903 | 0.838 | 0.934 | 0.998 | (0.570) | 0.900 | 0.999 | **0.933** |

Table 9: Pearson correlation between productivity and evaluation measures. Translator 8's results are omitted from the average.

### 4.4 Discussion

Analyzing the results further, there seems to be benefits of using FMS over BLEU or TER to evaluate MTPE. Table 10 shows the average BLEU, TER, FMS and productivity gain values for MTPE and the 75–84% fuzzy band. The average gain due to MTPE was higher than the 75–84% band. However, according to BLEU scores, the situation should have been the opposite,

while according to TER both values should have been more or less the same. The FMS, on the other hand, accurately reflects the productivity gains obtained.[14]

| | BLEU | TER | FMS | Prod. gain |
|---|---|---|---|---|
| **75–84%** | 70.70 | 20.98 | 85.60 | 22.52% |
| **MTPE** | 66.07 | 20.90 | 87.91 | 24.09% |

Table 10: Average BLEU, TER, FMS and productivity gain for MTPE and 75–84% bands.

Another useful application of FMS is the insight they provide when setting the threshold between TM matching and segments sent to the MT engine. Normally, TM matches below 75% are assumed to yield no productivity increase. Therefore, when applying MTPE, the general approach is to process all segments in the 0–74% range via MT. However, if the MT output quality is high enough, it may be more productive to post-edit MT output rather than some of the TM matches, as the results of this experiment actually show. This issue was first brought to our attention by the post-editors themselves, who frequently suggested increasing the threshold to 85%. After internal testing, it was verified that the post-editors' intuition was correct and a new threshold of 85% was established for this client.[15] Interestingly enough, that value is also the average FMS obtained with this client's projects over a three-year period. In our experience, this method can also be applied successfully to other language pairs and higher FMS ranges. For example, the threshold of Spanish-Catalan engines, which generally perform better than more distant language pairs, is set at 95%, as suggested by post-editors, confirmed by testing and reflected in FMS. These experiences suggest that the FMS applied to MTPE evaluation can be used to adjust the threshold between TM matching and MT segments on a client, domain or engine basis.

Since FMS seems to perform as well as other widely used metrics and incorporates other benefits, it may be reasonable to use FMS as a metric for analyzing both TM leverage and MT output. Relying on two different scores for these technologies creates contradictory situations, where a client may implicitly acknowledge that TM fuzzy matches below 75% do not yield productivity gains, and yet consider MT output with 60& FMS as liable for discounts based on their own assumptions on BLEU, TER or other metric. A unified framework for combining evaluation and quality estimation of TM and MT technologies has been recently proposed by Forcada and Sánchez-Martínez (2015). Even though they do not consider MT quality indicators and TM matching scores comparable, fuzzy scores still play an important role in this unification. In any case, these efforts may help to provide a better integration of translation technologies and spread best practices for all parties involved in a translation project.

## 5 Conclusion and future work

In this paper, we reported an experiment where ten professional translators with diverse experience in translation and MTPE complete the same translation and post-editing job within their everyday work environment using files from a real translation request. The objective was to analyze the relation between productivity and automated metrics in a commercial setting, and verify if the target-side FMS could be an adequate method of evaluating MT output for PE.

[14] This pattern has also been observed in other MTPE projects performed in our company.

[15] As explained in Section 3.1, for the purposes of the experiment reported here, this higher threshold was not used, so all segments below 75% were considered no-match segments. Also, we do not imply that MTPE can always be faster than post-editing 75–84% TM matches. With poor output, that threshold should still be set at 75% or even lower, while better engines may benefit from higher thresholds.

The more than 7,000 words of the file to be translated from English into Spanish was analyzed with the usual CAT tools in our company to ensure it contained a significant amount of fuzzy matches, exact matches and no-match segments. Half of the no-match segments was randomly selected for MTPE, while the other half was translated from scratch. The MT output for the MTPE sample was generated using one of our customized RBMT engines.

The results show that the FMS applied to MTPE evaluation has a strong correlation with productivity, comparable to the correlation obtained with more traditional evaluation metrics such as BLEU and TER. Moreover, the FMS was the only metric able to reflect the higher productivity obtained by the MTPE sample over the 75–84% TM match band. Another advantage shown is that it can also be used to set the TM matching threshold at which post-editing MT is more productive than post-editing TM matches.

Another finding is that MT output from a mature engine increases translators' productivity when compared to translation from scratch and can even surpass the throughput of TM matches. Only one of the ten translators involved did not experience any productivity gain. The productivity gain achieved by the rest of translators varied greatly between them, with the most experienced translator having the least productivity increase and the least experienced translator obtaining the biggest gain. However, this trend cannot be confirmed by the performance of the rest of translators.

It was also revealed that inline tags have a big impact on productivity, a fact which is not reflected in any of the known metrics and which has not yet received much attention in research. Other areas for future work include using this data set to perform a study on quality estimation to avoid depending on reference translations and further analysis of automated metrics in order to reveal the threshold at which each metric ceases to consider a segment useful in terms of productivity.

The performance of FMS applied to MTPE evaluation has been shown to be as solid as any of the traditional metrics. Its correlation with productivity is as strong as BLEU or TER and it can be used to determine thresholds between TM matching and MTPE segments. Moreover, FMS is easier to calculate for general users and we believe it would be more readily embraced by the industry due to its analogy with the well-established TM match bands.

## References


Aziz, W., de Sousa, S. C. M., and Specia, L. (2012). PET; a Tool for Post-Editing and Assessing Machine Translation. In *Eighth International Conference on Language Resources and Evaluation (LREC12)*, pages 3982–3987, Istanbul, Turkey. ELRA.

Bechara, H. (2013). Statistical Post-Editing and Quality Estimation for Machine Translation Systems. Master's thesis, Dublin City University, School of Computing.

Burchardt, A. and Lommel, A. (2014). Practical Guidelines for the Use of MQM in Scientific Research on Translation Quality. Technical report, DFKI, Berlin.

Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., and Schroeder, J. (2007). (Meta-) Evaluation of Machine Translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158.

Callison-Burch, C., Osborne, M., and Koehn, P. (2006). Re-evaluating the role of bleu in Machine translation research. In *Proceedings of the 11th Conference of the European Chapter of the Associatio for Computational Linguistics*, pages 249–256, Trento, Italy. EACL.

Denkowski, M. and Lavie, A. (2010). METEOR-NEXT and the METEOR paraphrase tables: improved evaluation support for five target languages. In *Proceedings of the Joint Fifth Workshop on Statistical*

*Machine Translation and MetricsMATR*, WMT '10, pages 339–342, Stroudsburg, PA, USA. Association for Computational Linguistics.

Dice, L. R. (1945). Measures of the Amount of Ecologic Association Between Species. *Ecological Society of America*, 26(3):297–302.

Doddington, G. (2002). Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, HLT '02, pages 138–145, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Federico, M., Cattelan, A., and Trombetti, M. (2012). Measuring User Productivity in Machine Translation Enhanced Computer Assisted Translation. In *Proceedings of the Tenth Conference of the Association for Machine Translation in the Americas (AMTA)*, San Diego, CA. AMTA.

Forcada, M. L. and Sánchez-Martínez, F. (2015). A general framework for minimizing translation effort: towards a principled combination of translation technologies in computer-aided translation. In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation*, pages 27–34, Antalya, Turkey.

Giménez, J. and Màrquez, L. (2010). Asiya: An Open Toolkit for Automatic Machine Translation (Meta-)Evaluation. *The Prague Bulletin of Mathematical Linguistics*, (94):77–86.

Hurtado Albir, A. (1995). La didáctica de la traducción. Evolución y estado actual. In Fernández, P. and Bravo, J., editors, *Perspectivas de la traducción*, pages 49–74. Universidad de Valladolid.

King, M., Popescu-Belis, A., and Hovy, E. (2003). FEMTI: Creating and Using a Framework for MT Evaluation. In *Proceedings of the Machine Translation Summit IX*, pages 224–231, New Orleans, LA.

Koehn, P. (2010). *Statistical Machine Translation*. Cambridge University Press.

Lavie, A. (2010). *Evaluating the Output of Machine Translation Systems*. AMTA, Denver, Colorado, USA.

Lavie, A. and Agarwal, A. (2005). METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of ACL Workshop on Intirnsic and Extrinsic Evaluation Measures for MT and/or Summarisation*, pages 65–72.

Lin, C.-Y. and Och, F. J. (2004). Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, ACL '04, Stroudsburg, PA, USA. Association for Computational Linguistics.

Lommel, A. (2015). Multidimensional Quality Metrics (MQM) Definition. Technical report, DFKI.

Melamed, I. D., Green, R., and Turian, J. P. (2003). Precision and recall of machine translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume of the Proceedings of HLT-NAACL 2003–short papers*, volume 2 of *NAACL-Short '03*, pages 61–63, Stroudsburg, PA, USA. Association for Computational Linguistics.

Moran, J., Saam, C., and Lewis, D. (2014). Towards desktop-based CAT tool instrumentation. In *Proceedings of the Third Workshop on Post-Editing Technology and Practice*, pages 99–112, Vancouver, BC. AMTA.

Nießen, S., Och, F. J., Leusch, G., and Ney, H. (2000). An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research. In *International Conference on Language Resources and Evaluation*, pages 39–45, Athens, Greece. European Language Resources Association (ELRA).

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2001). Bleu: a Method for Automatic Evaluation of Machine Translation. IBM Research Report RC22176 (W0109-022), IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, NY 10598.

Parra Escartín, C. and Arcedillo, M. (2015). A fuzzier approach to machine translation evaluation: A pilot study on post-editing productivity and automated metrics in commercial settings. In *Proceedings of the ACL 2015 Fourth Workshop on Hybrid Approaches to Translation (HyTra)*, pages 40–45, Beijing, China. ACL.

Plitt, M. and Masselot, F. (2010). A Productivity Test of Statistical Machine Translation Post-Editing in a Typical Localisation Context. *The Prague Bulletin of Mathematical Linguistics*, 93:7–16.

Ruopp, A. (2010). The Moses for Localisation Open Source Project. In *Proceedings of the Ninth Conference of the Association for Machine Translation in the Americas*, Denver, Colorado. AMTA.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, pages 223–231, Cambridge, Massachusetts, USA.

Snover, M., Madnani, N., Dorr, B. J., and Schwartz, R. (2009). Fluency, adequacy, or HTER? Exploring different human judgments with a tunable MT metric. In *In Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 259–268.

Sørensen, T. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. *Kongelige Danske Videnskabernes Selskab*, 5 (4):1–34.

Tillmann, C., Vogel, S., Ney, H., Zubiaga, A., and Sawaf, H. (1997). Accelerated DP Based Search For Statistical Translation. In *European Conf. on Speech Communication and Technology*, pages 2667–2670.

Zhechev, V. (2012). Machine Translation Infrastructure and Post-Editing Performance at Autodesk. In *AMTA 2012 Workshop on Post-Editing Technology and Practice (WPTP 2012)*, pages 87–96, San Diego, USA. Association for Machine Translation in the Americas (AMTA).