The Prague Bulletin of Mathematical Linguistics
NUMBER 108 JUNE 2017 343–354

# Questing for Quality Estimation A User Study

Carla Parra Escartín,[a] Hanna Béchara,[b] Constantin Orăsan[b]

[a] ADAPT Centre, SALIS, Dublin City University, Ireland
[b] RGCL, University of Wolverhampton, UK

**Abstract**

Post-Editing of Machine Translation (MT) has become a reality in professional translation workflows. In order to optimize the management of projects that use post-editing and avoid underpayments and mistrust from professional translators, effective tools to assess the quality of Machine Translation (MT) systems need to be put in place. One field of study that could address this problem is Machine Translation Quality Estimation (MTQE), which aims to determine the quality of MT without an existing reference. Accurate and reliable MTQE can help project managers and translators alike, as it would allow estimating more precisely the cost of post-editing projects in terms of time and adequate fares by discarding those segments that are not worth post-editing (PE) and have to be translated from scratch.

In this paper, we report on the results of an impact study which engages professional translators in PE tasks using MTQE. We measured translators' productivity in different scenarios: translating from scratch, post-editing without using MTQE, and post-editing using MTQE. Our results show that QE information, when accurate, improves post-editing efficiency.

## 1. Introduction

Machine Translation Post-Editing (MTPE) has become a reality in industrial translation settings. This has impacted on translation workflows as translators are imposed shorter deadlines and lower rates for these tasks than when translating from scratch. However, the quality of Machine Translation (MT) still remains an issue, particularly for post-editors, who usually complain that they spend more time assessing the MT output quality and fixing the translations, than when translating the same text from scratch. Many professional translators acknowledge that after a few segments un-

Corresponding author: carla.parra@adaptcentre.ie

Cite as: Carla Parra Escartín, Hanna Béchara, Constantin Orăsan. Questing for Quality Estimation A User Study. The Prague Bulletin of Mathematical Linguistics No. 108, 2017, pp. 343–354. doi: 10.1515/pralin-2017-0032.

dergoing MTPE, they delete the remaining segments and translate everything from scratch if they deem that it will take them less time. This suggests that in some cases the translations suggested are not good enough. MT Quality Estimation (MTQE) can address this issue by assessing the quality of an automatically translated segment and proposing for post-editing only those that are good enough.

Quality estimation in MT aims to predict the quality of the MT output without using a reference translation (Blatz et al., 2004; Specia et al., 2011). This field has received extensive interest from the research community in recent years, resulting in the proposal of a number of machine learning methods that estimate the quality of a translation on well defined data sets, but which do not necessarily reflect the reality of professional translators. In order to integrate MTQE successfully in translation workflows it is necessary to know when a segment is good enough for a translator. However, and as pointed out by Turchi et al. (2015), “QE research has not been followed by conclusive results that demonstrate whether the use of quality labels can actually lead to noticeable productivity gains in the CAT framework”.

In this paper, we present a user study which aims to understand better how quality estimation should be used in order to improve the productivity of professional translators. To achieve this, we use the English to Spanish part of the Autodesk Post-Editing data corpus (ISLRN 290-859-676-529-5) and 4 professional translators. As this data set comprises real data used by Autodesk in past translation projects, it constitutes a valid and publicly open dataset for our experiments to validate the usability of MTQE in real translation scenarios.

The remainder of this paper is structured as follows: in the next Section 2, we briefly discuss previous relevant work. Section 3 reports on the experimental setting on the study that we carried out, followed by analysis and discussion of the results in Section 4. Finally, Section 5 wraps up our work and discusses future paths to be explored.

## 2. Related Work

Although MTQE has not been widely tested in real translation workflows, a few researchers, particularly in the field of translation studies, have attempted to cover this gap and assess to which extent MTQE could be useful for professional translators. In their work, Turchi et al. (2015) assess whether the use of quality labels can actually lead to noticeable productivity gains. They do so by first establishing the conditions to carry out on-field evaluation and then carrying out an experiment providing translators with binary quality labels (green and red, depending on the MTQE obtained for the segment). They observed a non-significant productivity increase in translators’ productivity though. When dividing the test data according to segment length and quality, they concluded that “the higher percentage of wins is statistically significant only for medium-length suggestions with HTER>0.1”. Their data set accounted for 1389 segments (542 were used in training the QE engine, and 847 in testing) and their experiment was carried out by four professional translators. In total, they gathered two instances of each segment, one for the scenario in which the translator was shown the MT output QE, and one in which the translator did not have a QE of the MT output.

Moorkens et al. (2015) researched whether human estimates of post-editing effort accurately predict actual post-editing effort and whether the display of confidence scores (MTQE) influences post-editing behaviour. In their study, they used two different groups of participants. One consisting of six members of staff, postdoctoral researchers and PhD students of a Brazilian University, and a second one consisting of 33 undergraduate and Masters translation students. The first group of participants were asked to assess the quality of a set of 80 segments of two Wikipedia articles describing Paraguay and Bolivia and Machine Translated into Portuguese using Microsoft's Bing Translator. They were asked to classify the MT output according to a 3-grade scale:

1. Segments requiring a complete retranslation;
2. Segments requiring some post-editing but for which PE is still quicker than retranslation; and
3. Segments requiring little or no post-editing.

Secondly, and after a break of at least 2 weeks to avoid the participants remembering their ratings, the same participants were asked to post-edit the segments without any type of MTQE being shown. Finally, the second group of participants (undergraduate and masters students), were asked to post-edit the same sample but in this case MTQE was used. Using the average ratings of the first phase of their research, Moorkens et al. (2015) colour-coded each segment in red (better to retranslate), amber (MT could be useful but requires post-editing), and green (MT requires little or no post-editing). Although their study is based in a rather small sample, their findings suggest that "the presentation of post-editing effort indicators in the user interface appears not to impact on actual post-editing effort".

Moorkens and Way (2016) researched the acceptability of translation memory (TM) compared to that of MT among translators. They engaged 7 translators and asked them to rate 60 segments translated from English into German. The text was taken from the documentation of the an open-source computer-aided design program called FreeCAD and from the Wikipedia entry describing what computer-aid design is. They conclude that when low- and mid-ranking fuzzy matches are presented to translators without scores, translators find the suggestions irritating, and for over 36% of such instances, useless for their purposes. In contrast, in their experiment all of the MT matches suggested were rated as having some utility to post-editors. Moorkens and Way (2016) conclude that their findings suggest that "MT confidence measures need to be developed as a matter of urgency, which can be used by post-editors to wrest control over what MT outputs they wish to see, and perhaps more importantly still, which ones should be withheld".

Finally, in a recent study aiming at determining the user interface needs for post-editors of MT, Moorkens and O'Brien (2017) report that of the respondents to a survey aiming at determining the features that translators wished post-editing interfaces had, 81% expressed the wish of having a feature showing confidence scores for each target text segment from the MT engine. This finding makes the impact study reported here of utmost relevance, as we precisely aimed at investigating the impact of showing MTQE to translators when undergoing MTQE tasks. This will be explained in the next section 3.

## 3. Experimental Setup

### 3.1. Data

We decided to use the Autodesk Post-Editing Data corpus in our experiments in order to simulate a real translating experience. This corpus consists of parallel data with English as the source language and 12 different target languages. The size per language pair varies from 30,000 to 410,000 segments, and each segment is labelled with information as to whether it comes from a Translation Memory (TM) match or it is MT output. The post-edited target sentences are also included in the dataset, along with a raw MT score and a Fuzzy Match Score. The data belongs to a technical domain, and the segments come predominantly from software manuals.

We selected our sentences from the English to Spanish part of the corpus. We then used a semantically enriched version of QuEst++ (cf. Section 3.2) in order to predict the target-side Fuzzy Match Score (FMS) of the machine translation output. We decided to use the FMS as translators are more used to working with Translation Memory leveraging and fuzzy matches (Parra Escartín and Arcedillo, 2015a,b) than to more traditional MT evaluation metrics such as BLEU (Papineni et al., 2002) or HTER (Snover et al., 2006). While FMS is usually computed on the source side of a text, in our case, and similarly to what is proposed in Parra Escartín and Arcedillo (2015a,b), we use the FMS as a MT evaluation metric and thus aim at predicting a target-side FMS. Following the findings in Parra Escartín and Arcedillo (2015b), we established a threshold of 75% FMS or higher to consider a segment worth to be post-edited.

The sentences used in this experiment were selected in such a way that a quarter were chosen from sentences where the QE system performed well ("Good QE"). In these cases the predicted FMS for each sentence is close enough to the observed FMS score to give the translator a correct idea of its quality (within 5%). Another quarter were chosen from sentences where the QE system performed badly ("Bad QE"). In this case the observed score is more than 10% off compared to the observed score.[1] Another quarter of the sentences do not include MTQE information, and the final

[1] Given the difference between the predicted QE and the observed score some of the segments are being mislabelled as "worth post-editing"/"not-worth postediting". We took this into consideration during our evaluation.

quarter have no MT suggestion at all ("Translate From Scratch"). For the purpose of this study, our total number of sentences is 300 (about 3000 words) equally distributed among the four categories above (i.e. each category contained a total of 75 segments).

### 3.2. The Quality Estimation System

In our experiments, we use QuEst++ (Specia et al., 2015) enhanced with the semantically motivated features we described in (Béchara et al., 2016). QuEst++ is considered to be the state-of-the-art framework for MTQE tasks and is used as a baseline in the most recent MTQE shared tasks, such as the ones in 2014 (Bojar et al., 2014), 2015 (Bojar et al., 2015), and 2016 (Bojar et al., 2016). It includes a feature extraction framework and also provides with the machine learning algorithms necessary to build the MTQE prediction models. The 17 baseline features are language independent and include shallow surface features (e.g. number of punctuation marks, average length of words, number of words, etc.). They also include n-gram frequencies and language model probabilities.

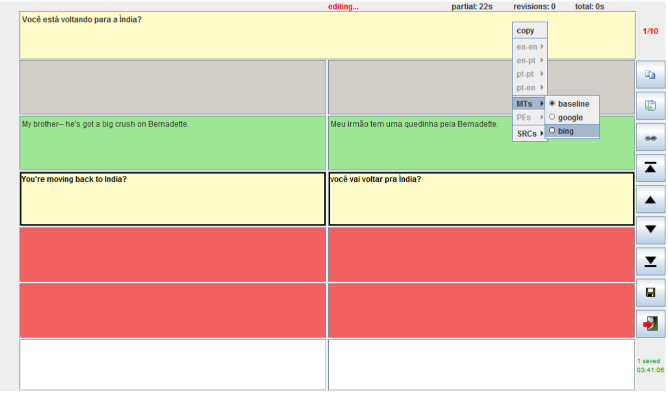
We tuned QuEst++ with in-domain data, building our own language models and n-gram counts from the Autodesk Dataset. As estated above, we also added a number of additional features to the system. More concretely, we extracted a variety of linguistically motivated features inspired by deep semantics such as distributional Similarity Measures, Conceptual Similarity Measures, Semantic Similarity Measures and Corpus Pattern Analysis (Béchara et al., 2016). We integrated these Semantic Textual Similarity (STS) features into the QE pipeline and noticed an improvement over the baseline. By replicating the experiments in Béchara et al. (2016) for the Autodesk data, we observe similar results as demonstrated in Table 1.

| **System Description** | MAE |
|---|---|
| QuEst++ – out of the box | 9.82 |
| QuEst++ – tuned for in-domain data | 9.78 |
| QuEst++ – with STS features | 9.52 |

*Table 1: MAE predicting the FMS for Autodesk*

### 3.3. PET: Post-Editing Tool

For our study we use PET (Aziz et al., 2012) as our post-editing tool. Like other CAT tools, PET provides an easy to use user interface which facilitates both translating and post-editing. In addition, the tool records a number of statistics such as the keystrokes pressed and the time needed to perform the translation, which are very relevant for this research. Even though PET is unlikely to be used in a real-world post-editing situation, it is ideal for our research. The tool is open-source and written in Java, which allowed us to easily modify the code to incorporate the traffic light system described in section 3.4. While other tools such as SDL Trados Studio[2] or MemoQ[3] would have been preferred by the translators due to both familiarity and ease of use, these tools did not allow the same kind of malleability and customisation as PET, which allowed us access to the source code in order to edit in our traffic lights.



Você está voltando para a Índia?

My brother-- he's got a big crush on Bernadette.

Meu irmão tem uma quedinha pela Bernadette.

You're moving back to India?

você vai voltar pra Índia?

Figure 1: A Screenshot of PET out of the box

### 3.4. The User Study

Inspired by the work reported by Turchi et al. (2015), we modified PET to present translators with a traffic light system which suggests the type of task they were facing in each case:

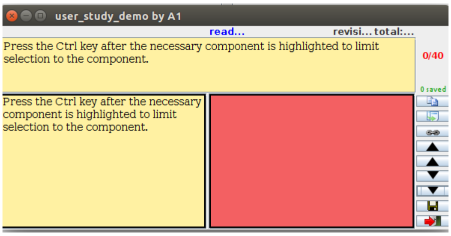
**Light yellow** (referred to in the evaluation as *Translate*) indicated that a translator had to translate the given sentence from scratch (in this case, the translator was not given an MT translated sentence to post-edit).

**Light blue** (*Post-edit*) indicated that a machine translation of the source segment is available, however, no MTQE information is provided, and therefore the translators must decide for themselves whether to translate from scratch, or to post-edit.

**Light green** (*QE Post-edit*) indicated that the MTQE system strongly suggests that the translator should post-edit the sentence produced by the MT engine. As indi-

[2] `http://www.sdl.com/solution/language/translation-productivity/trados-studio/`

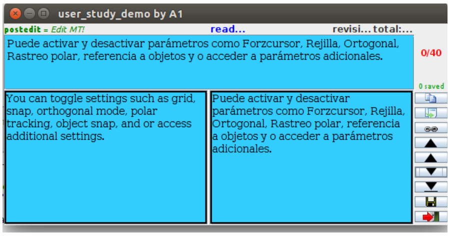
[3] `https://www.memoq.com/en/`



Press the Ctrl key after the necessary component is highlighted to limit selection to the component.

Press the Ctrl key after the necessary component is highlighted to limit selection to the component.

Figure 2: Translate from scratch



Puede activar y desactivar parámetros como Forzcursor, Rejilla, Ortogonal, Rastreo polar, referencia a objetos y o acceder a parámetros adicionales.

You can toggle settings such as grid, snap, orthogonal mode, polar tracking, object snap, and or access additional settings.

Puede activar y desactivar parámetros como Forzcursor, Rejilla, Ortogonal, Rastreo polar, referencia a objetos y o acceder a parámetros adicionales.

Figure 3: Post-edit without MTQE

cated earlier, this means that the MTQE system has predicted a fuzzy match score of 75% or more.

**Light red** (*QE Translate*) indicated that the MTQE system strongly suggests that the translator translates the sentence from scratch. This means that the MTQE system has predicted a fuzzy match score of less than 75%.

Figures 2 and 3 show how the colour coding system was displayed to the translators.

In order to refine our experiment, we performed a pilot study engaging 4 non-professional translators who are native speakers of Spanish. These translators were asked to look at a subset of 40 sentences from the full dataset. While the results of this study remained inconclusive in terms of linking productivity to MTQE, we learned a lot about the needs of the translators and the presentation of the task. For the full study, we enlisted the help of 4 Spanish professional translators with several years' translating experience. The years of experience varied greatly, between 3 and 14 years experience. All 4 translators had some experience with Computer-Assisted Translation tools and Post-Editing tasks. All 4 translators are native speakers of Spanish with a working proficiency of English and were asked to fill out questionnaires before and after completing the tasks with the aim of gathering information about their background and their experience while performing the task. Table 2 summarises the translator details.

While all translators had some experience with post-editing tools, none of the translators were familiar with PET before participating in the experiment. To overcome this issue, together with the instructions to carry out the task for the experiment, we also provided them with a short user manual of the tool with screenshots aiming at familiarising them with the interface prior to the task itself. All translators were paid for their time and were asked to complete the task over the course of a day, in order to simulate the real-world experience.

| Translator | C | M | V | S |
|---|---|---|---|---|
| Experience in technical domains (years) | 14 | 6 | 3 | 6 |
| Experience as a professional translator (years) | 14 | 6 | 3 | 9 |
| Experience with post-editing tools (years) | 2 | 4 | 3 | 1 |
| Opinion of Computer-Assisted Translation tools | Pos | Pos | Pos | Pos |
| Opinion of post-editing tasks | Neg | Pos | Pos | Pos |

*Table 2: Overview of the professional translators engaged in the experiment*

## 4. Results and Discussion

We extracted the post-editing times and keystrokes for all 4 translators. We then normalised these results by dividing each by the number of tokens in the final post-edited target sentence in order to compare sentences of different lengths. We also discarded one sentence, because the post-editing time exceeded 9000 seconds. In cases where a translator skipped a sentence, we discarded their statistics as well. In both cases we discarded the sentence data for all translators, in order to ensure the results remained comparable. In total, we discarded 4 sentences this way. In this section, we summarise the results on the remaining data.

Figure 4 shows the time, measured in seconds per word, that each translator spent on a given type of task (raw post-editing, translating from scratch, QE Postedit and QE translate). Each translator is identified by a letter. In addition, we provide the average for all four translators. As we expected, the sentences that needed to be translated from scratch took the most time across all translators, even without taking into account the quality of the QE. This seems to suggest that MT can considerably boost translator efficiency.

In Figure 5 we look closer at the time spent post-editing, separating out the good and bad QE. Here we can see that good quality estimation results, on the other hand, seem to consistently enhance performance across all translators. The average time spent per token drops from 1.62 seconds for no QE to 1.15 seconds for good QE.

In order to gain more insight, we also take a look at the number of keystrokes by type of task and by MTQE quality. Figures 6 and 7 take a closer look based on the type of task and the quality of MTQE respectively. The number of keystrokes used in post-editing is clearly lower than the number of keystrokes used when translating from scratch. Strangely enough, translators used less keystrokes in the cases where, despite being given a translation, they were instructed to translate from scratch, than when they were asked to post-edit the translation. This is an unexpected result and it will have to be investigated further. One possible explanation could be that they used the arrow keys a lot to navigate in the segment. The average number of keystrokes per segment drops from 81 for no QE to 46 with bad QE and 33 with good QE. Here we can see that even bad QE seems to be a better aid than no QE at all, at least in terms of

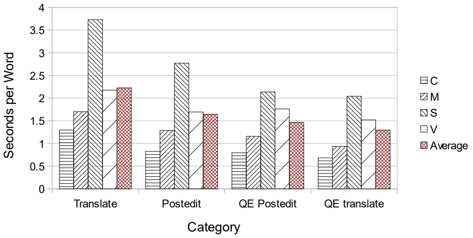


*Figure 4: Number of seconds per word spent translating/ post-editing per type of task*

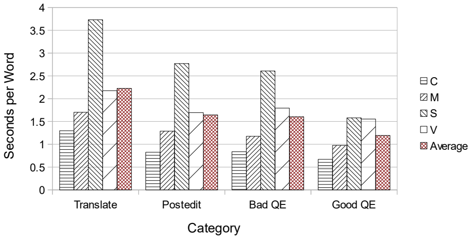


*Figure 5: Number of seconds per word spent translating/ post-editing by QE quality*

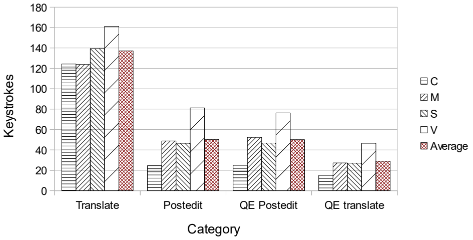


*Figure 6: Number of keystrokes per segment spent translating/ post-editing per type of task*

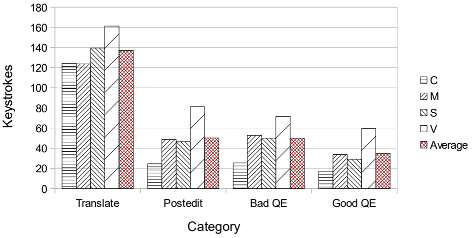


*Figure 7: Number of keystrokes per segment spent translating/ post-editing by QE quality*

post-editing effort as measured by keystrokes, and in this experimental setting. This might be because even the segments which are marked as "translate from scratch" provide a MT output which gives translators at least something to work with rather than starting from nothing.

As part of the experiment, we also asked all the translators to fill out questionnaires before and after the task in order to gain a more first-hand perspective of translators and post-editing tools. Responses suggest that while all four translators approved of the MT suggestions, all found the post-editing tool difficult to navigate, which may have affected both their results and opinions of MTQE. Despite our findings, three of the translators answered that they did not find MTQE helpful. However, as the translators had no way of distinguishing which was good and which was bad QE, this could have influenced their opinions of the usefulness of it. One translator disagreed, saying that they liked getting a first impression via the traffic lights system. Three out of the four translators claimed that the MT suggestions were helpful, while one insisted that they were better off translating from scratch, despite the high increase in efficiency shown by the results above.

## 5. Conclusion and Future Work

In this paper, we have reported on the results of a user study we conducted to investigate the impact of using the MTQE information in the post-editing workflow. We engaged 4 professional Spanish translators to take part in a post-editing/translation task, using a traffic lights system to provide MTQE information. We ran a study using 300 sentences from the Autodesk post-editing parallel corpus, annotated for Fuzzy Match Scores (FMS) using a semantically enhanced version of QuEst++.

Despite our rather small sample, our results seem to indicate that MTQE, especially good and accurate MTQE, is vital to the efficiency of the translation workflow, and can cut translating time and effort significantly. Translator feedback still seems quite negative in spite of this improvement, which suggests a better post-editing tool might be required to win over the translators.

In future work, we plan to analyse the results of this user study further. The data compiled through this experiment will also be released to allow other researchers to replicate our work or carry out further studies and/or experiments. We would also like to test whether the results reported here replicate for other language pairs and domains. Similar findings in such experiments would demonstrate the need for accurate and reliable MTQE, as well as the need to integrate it in professional translation workflows to improve post-editing efficiency. Our results, despite preliminary, seem to indicate this.

## Acknowledgements

The authors wish to thank the anonymous reviewers for their valuable feedback. This research is supported by the People Programme (Marie Curie Actions) of the European Union's Framework Programme (FP7/2007-2013) under REA grant agreement nº 317471, the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement Nº 713567, and Science Foundation Ireland in the ADAPT Centre (Grant 13/RC/2106) (www.adaptcentre.ie) at Dublin City University.

## Bibliography

Aziz, Wilker, Sheila Castilho, and Lucia Specia. PET: a Tool for Post-editing and Assessing Machine Translation. In *LREC*, pages 3982–3987, 2012.

Béchara, Hanna, Carla Parra Escartín, Constantin Orăsan, and Lucia Specia. Semantic Textual Similarity in Quality Estimation. *Baltic Journal of Modern Computing*, 4(2):256, 2016.

Blatz, John, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. Confidence estimation for machine translation. In *Proceedings of the 20th International Conference on Computational Linguistics (CoLing-2004)*, pages 315–321, 2004.

Bojar, Ondřej, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. Findings of the 2014 Workshop on Statistical Machine Translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA, June 2014. Association for Computational Lingustics.

Bojar, Ondřej, Rajan Chatterjee, Christian Federmann, Barry Haddow, Chris Hokamp, Matthias Huck, Varvara Logacheva, and Pavel Pecina, editors. *Proceedings of the Tenth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Lisbon, Portugal, September 2015.

Bojar, Ondřej, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. Findings of the 2016 Conference on Machine Translation. In *Proceedings of the First Conference on Machine Translation*, pages 131–198, Berlin, Germany, August 2016. Association for Computational Linguistics.

Moorkens, Joss and Sharon O'Brien. *Human Issues in Translation Technology: The IATIS Yearbook*, chapter Assessing User Interface Needs of Post-Editors of Machine Translation, pages 109–130. Routledge, Oxford, UK, 2017.

Moorkens, Joss and Andy Way. Comparing Translator Acceptability of TM and SMT outputs. *Baltic Journal of Modern Computing*, 4(2):141–151, 2016.

Moorkens, Joss, Sharon O'Brien, Igor A.L. da Silva, Norma B. de Lima Fonseca, and Fabio Alves. Correlations of perceived post-editing effort with measurements of actual effort. *Machine Translation*, 29(3–4):267–284, 2015.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 311–318, 2002.

Parra Escartín, Carla and Manuel Arcedillo. Machine translation evaluation made fuzzier: A study on post-editing productivity and evaluation metrics in commercial settings. In *Proceedings of the MT Summit XV*, Miami (Florida), October 2015a. International Association for Machine Translation (IAMT).

Parra Escartín, Carla and Manuel Arcedillo. Living on the edge: productivity gain thresholds in machine translation evaluation metrics. In *Proceedings of the Fourth Workshop on Post-editing Technology and Practice*, pages 46–56, Miami, Florida (USA), November 2015b. Association for Machine Translation in the Americas (AMTA).

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and Makhoul John. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of Association for Machine Translation in the Americas (AMTA)*, pages 223–231, 2006.

Specia, Lucia, Najeh Hajlaoui, Catalina Hallett, and Wilker Aziz. Predicting Machine Translation Adequacy. In *Proceedings of the 13th Machine Translation Summit*, pages 513–520, Xiamen, China, September 2011.

Specia, Lucia, Gustavo Paetzold, and Carolina Scarton. Multi-level Translation Quality Prediction with QuEst++. In *Proceedings of ACL-IJCNLP 2015 System Demonstrations*, pages 115–120, Beijing, China, July 2015. Association for Computational Linguistics and The Asian Federation of Natural Language Processing.

Turchi, Marco, Matteo Negri, and Marcello Federico. MT Quality Estimation for Computer-assisted Translation: Does it Really Help? In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers)*, pages 530–535, Beijing, China, July 26–31 2015. Association for Computational Linguistics.

**Address for correspondence:**
Carla Parra Escartín
carla.parra@adaptcentre.ie
ADAPT Centre, School of Applied Language and Intercultural Studies
Dublin City University
Glasnevin
Dublin 9, Ireland