Machine Translation Quality and Post-Editor Productivity

Marina Sanchez-Torron msnc017@aucklanduni.ac.nz
School of Cultures, Languages and Linguistics
University of Auckland
Auckland 1010, New Zealand
Philipp Koehn phi@jhu.edu
Department of Computer Science
Johns Hopkins University
Baltimore, MD 21218-2608, USA

Abstract

We assessed how different machine translation (MT) systems affect the post-editing (PE) process and product of professional English–Spanish translators. Our model found that for each 1-point increase in BLEU, there is a PE time decrease of 0.16 seconds per word, about 3-4%. The MT system with the lowest BLEU score produced the output that was post-edited to the lowest quality and with the highest PE effort, measured both in HTER and actual PE operations.

1 Introduction and Related Work

There is a relatively fair amount of empirical research on post-editing machine translation output (PE) focusing on assessing potential time and quality improvements over human translation with or without translation memories (TM). A common finding is that, despite differences among participants, PE is on average faster than unassisted or TM-assisted translation. Some examples of studies finding such speed benefits are those by Guerberof (2009), Flournoy and Duran (2009), Groves and Schmidtke (2009), Plitt and Masselot (2010) and Skadiņš et al. (2011).

In terms of PE quality, studies have shown, through the use of human judgments, that PE leads to quality comparable to (García, 2010) or better than other types of translation (Guerberof, 2009; Fiederer and O'Brien, 2009; Carl et al., 2011; Green et al., 2013; Läubli et al., 2013). There is therefore strong empirical evidence pointing at the speed and quality benefits of PE.

While many factors may affect productivity in a PE workflow, MT quality is one that is relatively easy to measure. Previous studies to have investigated how MT quality affects PE speed are those by Tatsumi (2009) and O'Brien (2011). They found different levels of correlations between MT quality, measured on a sentence level with automatic metrics, and PE speed. Krings (2001) and De Sutter (2012) too found that human judgments of MT quality correlated to PE speed. Koehn and Germann (2014) found their worst MT system entailed 20% more editing activity than their best one.

To the best of our knowledge, this is the first study to attempt to assess how different MT systems, of the same type but with different quality levels, affect the PE productivity of professional translators. By investigating how MT quality affects both PE time and PE quality, we aim at providing MT users and researchers with another approach to examining the usefulness of MT for PE purposes.

2 Study Setup

Nine translators were hired through ProZ[1], self-described as the biggest online translation workplace. Selected participants were offered a fixed compensation based on the standard, general, English–Spanish translation rates displayed on the website. Each translator post-edited four news texts of about 650 words each. Texts had similar complexity levels and were presented to translators in randomized order to dilute possible familiarization or fatigue effects. All four texts were translated into Spanish with nine MT systems (cf. Section 4.1). The output of all nine systems was assigned randomly to participants. As in Cettolo et al. (2013) and Koehn and Germann (2014), to deal with between-participant variability, the following restrictions were implemented:

- All translators post-edited all source sentences.
- No translator post-edited the same source sentence twice.
- All translators were exposed to roughly the same amount of output of all MT systems.

Translators were asked to post-edit to full, human-like quality. They worked remotely on the open-source, web-based, computer aided translation (CAT) tool CASMACAT (Cognitive Analysis and Statistical Methods for Advanced Computer Aided Translation)[2] (Alabau et al., 2013).

3 Translator Profile

All participants were native Spanish, professional translators, with at least 2 years' experience using CAT tools. All were educated to college level. Except for TR4 and TR5, all participants had degrees in Translation. Table 1 summarizes their main characteristics.

Translator	Translation experience (years)	PE certification	PE experience (years)
TR1	2 to 5	No	<2
TR2	2 to 5	No	2 to 5
TR3	5 to 10	Yes	5 to 10
TR4	>10	No	2 to 5
TR5	5 to 10	No	None
TR6	>10	Yes	<2
TR7	5 to 10	No	5 to 10
TR8	2 to 5	No	2 to 5
TR9	>10	No	2 to 5

Table 1: Translators' background

Translators' perceptions of PE and MT were elicited through eight questions answered with a Likert scale. Because of the small sample size, we grouped the answers from the five initial levels into three: *Strongly disagree* and *Disagree* were grouped into a new level *No*; and *Agree* and *Strongly agree* into a new level *Yes*. *Neutral* was left as such. Table 2 displays a summary of participants' PE and MT perceptions.

[1] http://www.proz.com
[2] http://www.casmacat.eu/

	No	Neutral	Yes
I am comfortable post-editing to human-like (perfect) quality	3	1	5
I am comfortable post-editing to less-than-perfect quality	4	1	4
I prefer PE to translating from scratch (without a TM)	3	2	4
MT helps me maintain translation consistency	4	1	4
MT helps me translate faster	2	4	3
PE is more laborious than translating from scratch or with a TM	3	4	2
I prefer PE to processing 85-94% TM matches	1	5	3
I prefer PE to editing a human translation	5	3	1

Table 2: Translators' perceptions of PE and MT

Answers show therefore a mix of opinions towards PE. Overall, translators are comfortable post-editing but do not prefer it to editing a human translation.

4 Variables of interest

4.1 Machine Translation Quality

Initiatives providing free resources for MT development and evaluation help boost MT research collaboration and efforts. One of these efforts is the WMT evaluation campaign of the Association for Computational Linguistics (ACL). Among the data they freely provide are training and test data sets. As part of their 8th Workshop on Statistical Machine Translation (WMT 2013)[3], the organizers released a test set comprised of 3,000 source English sentences and their corresponding Spanish human reference translations.

We trained nine MT systems with training data from the European Parliament proceedings, News Commentary, Common Crawl, and United Nations. The systems are phrase-based Moses systems (Koehn et al., 2007) with hierarchical lexicalized reordering (Galley and Manning, 2008), operation sequence model (Durrani et al., 2013), and sparse lexical features, using all available language model data (target side of the full parallel corpus, plus the provided monolingual news corpus and LDC Gigaword). The best system reaches comparable quality to the best system participating in the WMT 2013 evaluation campaign.

We iteratively halved the parallel training corpus to obtain systems of inferior quality. The quality of the MT systems was measured with case-sensitive BLEU (Papineni et al., 2002) on the official WMT 2013 test set. Table 3 summarizes MT systems' quality and training corpus size.

System	BLEU	Training sentences	Training words (English)
MT1	30.37	14,700k	385M
MT2	30.08	7,350k	192M
MT3	29.60	3,675k	96M
MT4	29.16	1,837k	48M
MT5	28.61	918k	24M
MT6	27.89	459k	12M
MT7	26.93	230k	6.0M
MT8	26.14	115k	3.0M
MT9	24.85	57k	1.5M

Table 3: MT Systems' quality and training corpus size

[3] http://www.statmt.org/wmt13/

4.2 Human-mediated Translation Edit Rate

Human-mediated Translation Edit Rate (HTER; Snover et al., 2006) measures the minimum number of operations (insertions, deletions, substitutions and shifts) needed to convert a machine translation into its post-edited version. It is computed by dividing the number of operations by the number of words in the post-edited version. HTER can be used as a measure of MT quality: the fewer the changes that need to be applied to the machine translation, the more similar it is to the post-edited, reference translation and therefore the higher the MT quality. Likewise, HTER can also be used as a measure of technical PE effort: the fewer changes necessary to convert the machine translation into its post-edited version, the less the effort exerted by the translator.

In this study, HTER scores were computed for all nine systems, based on the machine translated texts and their non-minimally post-edited versions, with the freely available tercom software[4].

4.3 Actual Edit Rate

HTER is concerned about the PE product, not the process. It therefore does not measure translators' actual edit operations, which may involve going back and applying corrections to previously post-edited parts of the text. We are interested in examining how actual edit operations vary across systems. Edit operations (i.e., insertions and deletions) are measured as the keystroke and/or mouse combinations leading to the insertion or deletion event, which do not necessarily correspond to the number of characters inserted or deleted. For instance, if a translator deletes a word by selecting it with the mouse and pressing *backspace*, it counts as one deletion event. If they move a word by cutting it and pasting it somewhere else, it constitutes one deletion and one insertion event. Insertion and deletion events were normalized by the number of words in the machine translated text to obtain a what we call here the *Actual Edit Rate*, henceforth AER.

4.4 Post-editing Time

Mean PE time per word per system was calculated by dividing the time spent post-editing each MT system's output by the number of source words translated by each system. Table 4 shows time measurements as PE time in seconds per word (spw) and as PE speed in words per hour (wph).

System	Mean PE time (spw)	Mean PE speed (wph)
MT1	4.06	887
MT2	4.38	822
MT3	4.23	851
MT4	4.54	793
MT5	4.35	828
MT6	4.36	826
MT7	4.66	773
MT8	4.94	729
MT9	5.03	716

Table 4: Systems' mean PE time (seconds per word) and PE speed (words per hour)

[4] https://github.com/jhclark/tercom

4.5 Post-editing Quality

Error-based quality assessment frameworks allow for the quantification of translation quality according to the type and severity of errors present in the translation. Previous PE studies (Guerberof, 2009; Temizöz, 2013) have used LISA's Quality Model 3.1, one such framework traditionally used in localization settings, to evaluate the quality of the post-edited texts. This model, however, is not officially available anymore. We therefore assessed PE quality with another error-based model, a quality metric compliant with QTLaunchPad's Multidimensional Quality Metric (MQM) framework.

We used the decision trees and guidelines provided in Burchardt and Lommel (2014) as a reference during the issue annotation process. Quality was assessed along the dimensions of Accuracy (Mistranslations, Omissions, Untranslated and Additions) and Fluency (Grammar, Style and Typography). We excluded Spelling from our quality metric as some translators pointed out that the spell checker did not work. Following LISA's standard weight scale, minor errors were given a weight of 1 and major errors were given a weight of 5.

5 Analysis and Results

Sentences with a logged period of inactivity of 2 minutes or more were excluded from analysis as such long pauses are likely not indicative of difficulties posed by the underlying MT system. Also excluded were sentences inadvertently skipped by translators and sentences with unreliable measurements[5]. In total, out of 1233 observations, 1202 were considered valid and submitted for analysis.

5.1 Post-Editing Time by Machine Translation System

For each MT system, we plotted mean PE time against BLEU score:

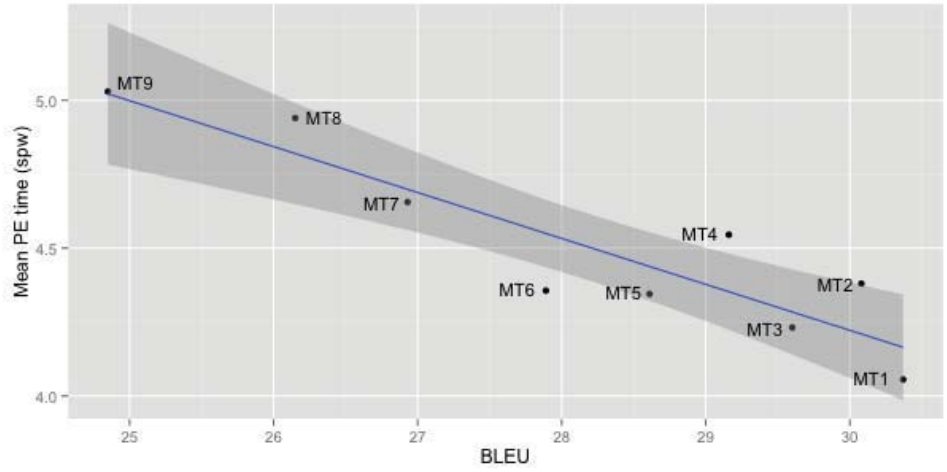

Figure 1: Scatter plot of systems' mean PE time against systems' BLEU and regression line with 95% confidence bounds

[5]CASMACAT logs editing time on a segment basis as the interval between the opening and closing of each segment. When translators access the PE interface, the first segment in the document is opened automatically. Translators do not usually start post-editing right away, instead they scroll through the document. Once acquainted with its contents, they go up the first segment, post-edit it and, when finished, close it. Most editing times logged for first segments in our study are in fact spent browsing through the whole document and are therefore unreliable.

A linear regression was applied to predict PE time based on MT quality. Our model was significant ($F(1,7) = 33.62$), with an R^2 of .828. It describes the effect of system's BLEU on mean PE time as following a decreasing linear relationship. Specifically, for every 1-point increase in BLEU, there is a decrease in PE time of approximately 0.16 seconds per word.

	Estimate	Std. error	*t* value	*p* value	95% confidence interval
intercept	8.88	0.76	11.74	<.001	[7.37, 10.39]
slope	-0.16	0.03	-5.80	<.001	[-0.21, -0.10]

Table 5: Linear regression results

Assumption checks confirm the validity of the results: the plot of residual versus fitted values shows some noise but no distinctive pattern, and although residuals show a slight departure from normality, this is expected in small samples.

5.2 HTER and AER by Machine Translation System

We investigate how both HTER and AER vary between MT systems. Table 6 displays HTER scores and mean AER by MT system:

System	HTER	AER
MT1	40.75	3.36
MT2	40.85	3.05
MT3	42.41	3.03
MT4	41.57	3.58
MT5	42.29	3.61
MT6	43.57	3.66
MT7	44.79	3.33
MT8	46.15	3.57
MT9	50.30	4.20

Table 6: Systems' HTER (%) and AER (events per word)

As expected, we see an almost continuous gradual increase in HTER as the quality of the MT system decreases. In contrast, our data does not allow us to establish any significant association between AER and MT quality. Keyboard activity may just not be as sensitive to MT quality as PE time. Nevertheless, MT9, the system with the lowest BLEU score has both the highest HTER and AER of all systems, representing an increase of 23.43% and 22% respectively over MT1, the best system.

5.3 Post-Editing Quality by Machine Translation System

Using an error-based framework to assess text quality usually involves determining an arbitrary pass/fail threshold for textual units. There is not a theoretical body of literature concerning these frameworks so we set the minimum sentence quality acceptance level at a percentage commonly referred to in industry documents, i.e., 95%[6].

PE quality is measured for the whole post-edited output, via both the MQM score and the count of sentences falling in the *Fail* and *Pass* categories, and reported by MT system. Also reported are normalized issue counts classified according to their severity (note that no critical errors were found in any of the texts).

[6] An error-free translation scores 100%. To calculate a sentence's MQM score with standard LISA severity weights the following formula applies: MQM Score (%) = 100 - (($Issues_{Minor}$ + 5 * $Issues_{Major}$ + 10 * $Issues_{Critical}$)/Sentence length)*100. A 28-word sentence with 2 minor issues and 1 major issue would have therefore a score of 100-(2+5)/28*100= 75%.

System	MQM Score	Fail	Pass	Minor issues/k	Major issues/k
MT1	97.86	15	117	12.39	1.72
MT2	97.19	20	113	12.83	3.12
MT3	98.01	16	117	12.59	1.75
MT4	96.76	26	109	15.32	3.40
MT5	97.84	18	116	11.93	2.04
MT6	98.63	10	124	9.88	1.02
MT7	97.31	17	115	11.39	3.45
MT8	97.47	22	112	10.48	3.38
MT9	95.81	32	103	14.28	4.76

Table 7: Indicators of translation quality of post-edited translations by MT system

As expected, given that participants are professional translators post-editing to high quality, sentence-level MQM scores follow a left-skewed distribution (867 of the 1202 sentences score 100%). A Pearson's chi-square test found differences in the Fail/Pass proportions between MT systems ($x^2(8)= 19.40$, $p < .05$), with a post-hoc pairwise comparison with Holm's adjustment finding significant the differences between the MT6-MT9 pair. While the Fail/Pass ratio or the MQM scores are not significantly different for all the other pairs, MT9, the system with the lowest BLEU score, produced the output that ended up with the lowest MQM score and the highest Fail/Pass ratio: 1 in 4 sentences were post-edited to below the acceptable 95% MQM score.

In terms of error categories, minor issues are more or less equally divided into Fluency and Adequacy issues across systems, while major issues are practically all Adequacy issues, mostly Mistranslations.

5.4 Post-Editing Quality vs. Post-Editing Time by Machine Translation System

We plotted the Fail/Pass ratio of the post-edited output against the mean PE time for each MT system:

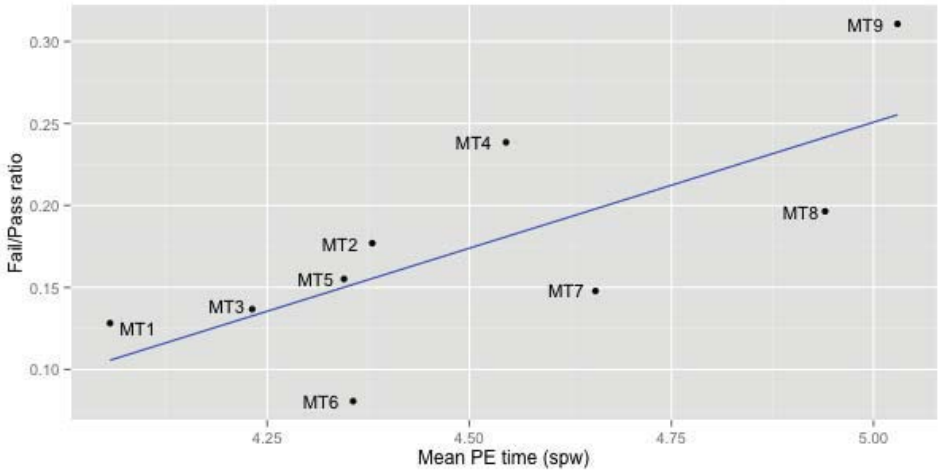

Figure 2: Scatter plot of systems Fail/Pass ratios against mean PE times with regression line

Our model describes a positive linear relationship between PE time and Fail/Pass ratio: the more time is spent post-editing, the higher the Fail/Pass ratio of the post-edited output. The

model is significant ($F(1,7) = 8.06$), with an R^2 of .535. Such a low R^2 points at additional factors affecting the quality of the post-edited texts. This will be investigated in further studies.

5.5 Post-Editing Time and Quality by Translators

Table 8 shows that differences in HTER, AER and PE time between translators are more pronounced than between MT systems.

	HTER	AER	Mean PE time (spw)	MQM Score	Fail	Pass
TR1	44.79	2.29	4.57	98.65	10	124
TR2	42.76	3.33	4.14	97.13	23	102
TR3	34.18	2.05	3.25	96.50	26	106
TR4	49.90	3.52	2.98	98.10	17	120
TR5	54.28	4.72	4.68	97.45	17	119
TR6	37.14	2.78	2.86	97.43	24	113
TR7	39.18	2.23	6.36	97.92	18	112
TR8	50.77	7.63	6.29	97.20	19	117
TR9	39.21	2.81	5.45	96.48	22	113

Table 8: Indicators of translation quality of post-edited translations by translator

The inter-subject variability reported in Table 8 mirrors the findings of previous empirical PE studies such as those mentioned in Section 1.

The slowest translator, TR7 has nevertheless both low HTER and AER. TR7's log shows they left the CASMACAT interface (by accessing another tab in the browser) and re-accessed CASMACAT an average of over 100 times per text. Without a screen recorder, we do not know what the translator was doing outside CASMACAT, but it is likely that they were engaged in translation-related web searches, possibly because the texts posed comparatively more difficulties for them.

TR8, the second slowest post-editor, has the second highest HTER and the highest AER. Comparing both variables among translators, we see that TR8 has a HTER comparable to that of TR4, yet TR8's AER is more than double that of TR4. This indicates that TR8 did, on average, considerable overwriting before settling on a final post-edited version, likely slowing them down in the process.

In our set, the two fastest translators, TR6 and TR4, left the CASMACAT interface the fewer number of times of all (5 and 3 times per text, respectively, on average), an indication that they did not need to consult many online translation resources. TR6 and TR4 are not only the fastest but also the most experienced translators of all participants, considering experience both in terms of length (>10 years for both) and translation volume in the preceding 12 months (40,000-55,000 and 25,000-39,900 words, respectively).

The two translators with industry PE certifications, TR6 and TR3, were the first and third fastest post-editors. They produced the texts with the two lowest HTER scores. While differences in the quality of the post-edited output are not statistically significant for translators, TR6 and TR3 produced two of the three translations with the highest Fail/Pass ratios. The second highest Fail/Pass ratio was produced by TR2, the less experienced translator, both in length of experience (2 to 5 years) and translation volume in the 12 preceding months (<10,000 words).

Lastly, we investigated the relationship between translators' PE time and PE quality by plotting Fail/Pass ratio against mean PE times. We did not find any association between translators' PE time and PE quality, as evidenced by the lack of pattern in the scatter plot in Figure 3:

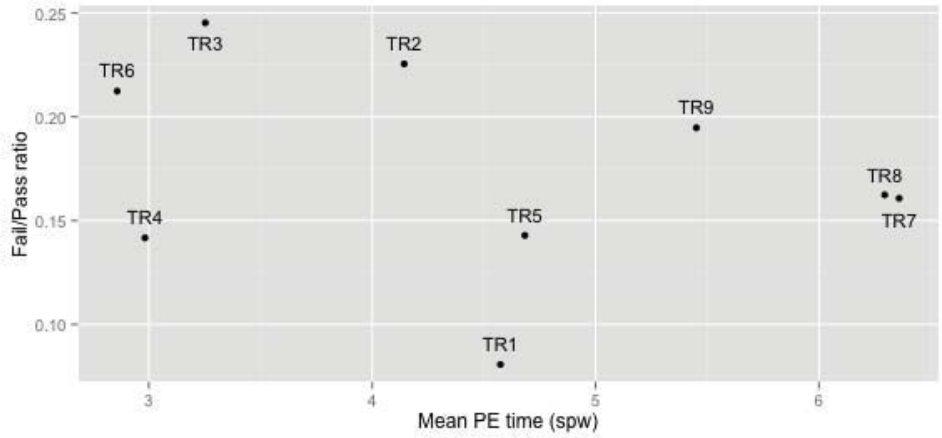

Figure 3: Scatter plot of translators Fail/Pass ratios against mean PE times with regression line

6 Conclusions

We presented a study that measured the impact of machine translation quality on post-editor speed and final translation quality. We found a linear relationship between machine translation quality, as measured by the BLEU score of the system, and post-editing speed of about 0.16 seconds/word post-editing time decrease per BLEU point increase. This is about a 3–4% speed increase for each BLEU point. We also found that worse machine translation output ultimately led to worse translation quality after post-editing. As future lines of research, we suggest investigating whether these findings extend to other language pairs.

Acknowledgements

This research was supported by a grant from The University of Auckland's Faculty of Arts Doctoral Research Fund.

References

Alabau, V., Bonk, R., Buck, C., Carl, M., Casacuberta, F., García-Martínez, M., González, J., Koehn, P., Leiva, L., Mesa-Lao, B., et al. (2013). CASMACAT: An open source workbench for advanced computer aided translation. *The Prague Bulletin of Mathematical Linguistics*, 100:101–112.

Burchardt, A. and Lommel, A. (2014). Practical guidelines for the use of MQM in scientific research on translation quality. `http://www.qt21.eu/downloads/MQM-usage-guidelines.pdf`. Accessed 07/12/2016.

Carl, M., Dragsted, B., Elming, J., Hardt, D., and Jakkobsen, A. L. (2011). The process of post-editing: a pilot study. In *8th International Conference NLPSC workshop. Special issue: Human-machine interaction in translation. Special theme: Human-machine interaction in translation*, pages 131–142.

Cettolo, M., Niehus, J., Stlker, S., Bentivogli, L., and Federico, M. (2013). Report on the 10th IWSLT evaluation campaign. In *10th Workshop on Spoken Language Translation (IWSLT)*.

De Sutter, N. (2012). MT evaluation based on post-editing: a proposal. In Depraetere, I., editor, *Text, Translation, Computational Processing [TTCP] : Perspectives on Translation Quality*, pages 125–144. De Gruyter Mouton, Berlin and Boston.

Durrani, N., Fraser, A., Schmid, H., Hoang, H., and Koehn, P. (2013). Can Markov models over minimal translation units help phrase-based SMT? In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Sofia, Bulgaria. Association for Computational Linguistics.

Fiederer, R. and O'Brien, S. (2009). Quality and machine translation: A realistic objective? *The journal of Specialised translation*, 11:52–72.

Flournoy, R. and Duran, C. (2009). Machine translation and document localization at Adobe: From pilot to production. In *Machine Translation Summit XII*.

Galley, M. and Manning, C. D. (2008). A simple and effective hierarchical phrase reordering model. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 848–856, Honolulu, Hawaii.

García, I. (2010). Is machine translation ready yet? *Target*, 22(2):7–21.

Green, S., Heer, J., and Manning, C. D. (2013). The efficacy of human post-editing for language translation. In *2013 IGCHI Conference on Human Factors in Computing Systems*, pages 439–448.

Groves, D. and Schmidtke, D. (2009). Identification and analysis of post-editing patterns for MT. In *Machine Translation Summit XII*, pages 429–436.

Guerberof, A. (2009). Productivity and quality in the post-editing of outputs from translation memories and machine translation. *Localisation Focus. The International Journal of Localisation*, 7(1):11–21.

Koehn, P. and Germann, U. (2014). The impact of machine translation quality on human post-editing. In *EACL 2014 Workshop on Humans and Computer assisted Translation*.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: open source toolkit for statistical machine translation. In *ACL*. Association for Computational Linguistics.

Krings, H. P. (2001). *Repairing Texts: Empirical Investigations of Machine Translation Post-Editing Processes*. Kent, Ohio, Kent State University Press.

Läubli, S., Fishel, M., Massey, G., Ehrensberger-Dow, M., and Volk, M. (2013). Assessing post-editing efficiency in a realistic translation environment. In *Workshop on Post-editing Technology and Practice*, pages 83–91.

O'Brien, S. (2011). Towards predicting post-editing productivity. *Machine Translation*, 25(3):197–215.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. 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.

Skadiņš, R., Puriņš, M., Skadiņa, I., and Vasiļjevs, A. (2011). Evaluation of SMT in localization to under-resourced inflected language. In *15th International Conference of the European Association for Machine Translation (EAMT)*, pages 35–40.

Snover, M., Dorr, B., Schwartz, R., Makhoul, J., Micciulla, L., and Weischedel, R. (2006). A study of translation error rate with targeted human annotation. In *7th Biennial Conference of the Association for Machine Translation in the Americas (AMTA 2006)*, pages 223–231.

Tatsumi, M. (2009). Correlation between automatic evaluation metric scores, post-editing speed, and some other factors. In *Machine Translation Summit XII*.

Temizöz, Ö. (2013). *Postediting Machine Translation Output and its Revision: Subject-matter experts versus Professional Translators*. PhD thesis, Universitat Rovira i Virgili.