Chapter 2

Comparing NMT and PBSMT for post-editing in-domain formal texts: A case study

Sergi Álvarez[a], Toni Badia[a] & Antoni Oliver[b]
[a]Universitat Pompeu Fabra [b]Universitat Oberta de Catalunya

This paper details a comparative analysis between phrase-based statistical machine translation (PBSMT) and neural machine translation (NMT) for English-Spanish in-domain medical documents using human rankings, fluency and adequacy, and post-editing (technical and temporal) effort, performed by professional translators. When MT output is ranked against translations performed by professional translators, results show a clear preference for human translations, with NMT in the second position. Regarding MT outputs, NMT is perceived as more fluent and conveying better the meaning of the source sentence. Despite this preference, post-editing temporal effort does not improve significantly in NMT compared to PBSMT, although technical effort is reduced.

1 Introduction

Over the last years, post-editing of machine translation (PEMT) has become common practice in the translation industry. It has been included as part of the translation workflow because it increases productivity and reduces costs (Guerberof 2009a). A recent survey showed that more than half of the language service providers (LSPs) offered PEMT as a service (Lommel & DePalma 2016). Post-editors "edit, modify and/or correct pre-translated text that has been processed by an MT system from a source language into (a) target language(s)" (Allen 2003).

Sergi Álvarez, Toni Badia & Antoni Oliver. 2021. Comparing NMT and PBSMT for post-editing in-domain formal texts: A case study. In Tra&Co Group (ed.), *Translation, interpreting, cognition: The way out of the box*, 33–46. Berlin: Language Science Press. DOI: 10.5281/zenodo.4545031

Yet, many professional translators state that after post-editing a few MT segments, they delete the remaining segments and translate everything from scratch if they consider it will take them less time (Parra Escartín & Arcedillo 2015).

Effective PE, therefore, requires sufficient quality of the MT output. The issue, then, is how to detect that a machine translation output is good enough to serve as input to PE. Very often, the usual automatic metrics do not always correlate to PE effort (Koponen 2016). Even translators' perception does not always match PE effort (Koponen 2012; Moorkens 2018). Research in this field has mainly focused on measuring the PE effort related to MT output quality (Guerberof 2009a,b; Specia 2011; 2010), productivity (O'Brien 2011; Parra Escartín & Arcedillo 2015; Plitt & Masselot 2010; Sanchez-Torron & Koehn 2016), translator's usability (Castilho et al. 2014; Moorkens & O'Brien 2013) and perceived PE effort (Moorkens et al. 2015).

Statistical machine translation (SMT) has been well established as the dominant approach in machine translation for many years. However, in the last few years, research has become more interested in neural machine translation after the computational limitations have been solved (Bahdanau et al. 2018; Cho et al. 2014). The first results obtained have been very successful in terms of quality, for example in WMT 2016 (Bojar et al. 2016), WMT 2017 (Bojar et al. 2017), and WMT 2018 (Bojar et al. 2018). These promising results have driven a technological shift from (phrase-based) statistical machine translation (SMT) to neural machine translation (NMT) in many translation industry scenarios.

All of the current research on post-editing machine translation output uses the division established by Krings (2001) regarding PE effort: temporal effort (time spent PE), technical effort (number of edits, often measured using keystroke analysis), and cognitive effort (usually measured with eye-tracking or think-aloud protocols). Even though no current measure includes all three dimensions, cognitive effort correlates with technical and temporal PE effort (Moorkens et al. 2015). In our experiments, we use automatic measures of both temporal and technical effort.

As this new approach to MT becomes more popular among LSPs and translators, it is essential to test what NMT can offer for PE in terms of quality compared to the results of PBSMT. Recent studies (Bentivogli et al. 2016; Castilho, Moorkens, Gaspari, Sennrich, et al. 2017; Toral & Sánchez-Cartagena 2017) have stated an improved quality of NMT for PE. In this paper, we continue in this direction, but we focus on in-domain formal documents, which are the ones usually post-edited by professional translators.

Our objectives with these experiments are threefold:

- Determine which MT method (PBSMT or NMT) yields better results for PE in-domain formal texts.
- Analyze the relation between human and automatic metrics for PE.
- Study translators perception as a prospective measure of PE effort.

In Section 2, we review previous work comparing SMT and NMT approaches. In Section 3 we describe the MT systems and the training corpus used. In Section 4 we include the automatic evaluation of the MT systems used. We give details about the methodology used for our experiments in Section 5. We explain the results obtained in Section 6 and, finally, we state the main conclusions and our plans for future work in Section 7.

2 Previous work

One of the first complete papers studying the impact of SMT and NMT in PE was Bentivogli et al. (2016). In it, they carry out a small scale study on post-editing NMT and SMT outputs of English to German translated TED talks. They conclude that NMT generally decreases the PE effort, but degrades faster than SMT with sentence length. One of the main strengths of NMT is the reodering of the target sentence.

Wu et al. (2016) evaluate the quality of NMT and SMT, in this case using BLEU (Papineni et al. 2002) and human scores for machine-translated Wikipedia entries. Results show that NMT systems outperform and improve the quality of MT results. Other studies have confirmed this diagnostics (Junczys-Dowmunt et al. 2016; Isabelle et al. 2017), as have the results of the automatic PE tasks at the Conference on Machine Translation (Bojar et al. 2016; 2017).

Toral & Sánchez-Cartagena (2017) broaden the scope of Bentivogli et al. (2016) adding different language combinations and metrics, and they conclude that although NMT yields better quality results in general, it is negatively affected by sentence length, and the improvement of the results is not always perceivable in all language pairs.

Castilho et al. (2017) discuss three studies using automatic and human evaluation methods. One of them includes in-domain formal texts for chemical patent titles and abstracts. In addition to the automatic metrics, two reviewers assess 100 random segments to rank the translations and to identify translation errors. Automatic evaluation doesn’t give clear results, but the SMT system is ranked higher than NMT in human evaluation.

Castilho et al. (2017) report on a comparative study of PBSMT and NMT, with four language pairs and different automatic metrics and human evaluation methods. It highlights some strengths and weaknesses of NMT, which in general yields better results. The study focuses especially on PE and uses the PET interface (Aziz et al. 2012) to compare educational domain output from both systems using different metrics. They conclude that NMT reduces word order errors and improves fluency for certain language pairs, so fewer segments require PE, especially because there is a reduction in the number of morphological errors. However, they don't detect a decrease in PE effort nor a clear improvement in omission and mistranslation errors.

Our experiments study the differences of post-editing NMT and SMT outputs for formal in-domain texts. We compare the usual automatic scores for MT with direct and indirect PE effort metrics. Mainly, we study translators' perception regarding quality, and fluency and accuracy, and analyze temporal and technical pot-editing effort.

3 MT systems and training corpus

3.1 MT systems

In order to help contextualise the results in our experiments, we have decided to use two MT systems as references to compare their results with the ones of the systems we trained. As reference MT systems, we have chosen Apertium (Forcada et al. 2011), a shallow transfer MT system, and Google Translate, a neural MT system for the English-Spanish language pair, which is the one we use in our experiments.

For training the PBSMT and neural MT systems we have used ModernMT (Germann et al. 2016) version 2.4. This version allows to train both statistical and neural MT systems. We have used the default options for this version. One of the salient characteristics of ModernMT is the fact that it can take into account the context of the sentence to be translated. In the evaluation results, we show figures for both cases: with and without taking the context into account. In the experiments we take context to be the previous and the next segment (except for the first and last segment, where we have taken into account the next and the previous segment only, respectively). Short contexts are usually enough to calculate the context vector used by ModernMT.

3.2 Data: Medical corpus

To train the system, we have compiled all of the publicly available corpora in the English-Spanish pair known to us. We have also created several corpora from websites with medical content:

- The EMEA[1] (*European Medicines Agency*) corpus.
- The IBECS[2] (*Spanish Bibliographical Index in Health Sciences*) corpus.
- Medline Plus:[3] we have compiled our own corpus from the web and we have combined this with the corpus compiled in MeSpEn[4].
- MSDManuals[5] English-Spanish corpus, compiled for this project under permission of the copyright holders.
- Portal Clínic[6] English-Spanish corpus, compiled by us for this project.
- The PubMed[7] corpus.
- The UFAL Medical Corpus[8] v1.0.

We have also treated as a corpus glossaries and glossary-like databases containing a lot of useful terms and expressions in the medical domain. Namely, we have used the English-Spanish glossary from MeSpEn, the 10th revision of the international statistical classification of ICD and SnowMedCT.

With all the corpora and glossaries we have created an in-domain training corpus of 2,836,580 segments and entries. We have split the corpus in two parts: 99% of the segments for training, and the remaining 1% for testing.

We have also used other general corpora for training the MT systems, namely the Scielo corpus, the Europarl corpus[9] (Koehn 2005), Global Voices corpus [10] and

[1] http://opus.nlpl.eu/EMEA.php

[2] http://ibecs.isciii.es

[3] https://medlineplus.gov/

[4] http://temu.bsc.es/mespen/

[5] https://www.msdmanuals.com/

[6] https://portal.hospitalclinic.org

[7] https://www.ncbi.nlm.nih.gov/pubmed/

[8] https://ufal.mff.cuni.cz/ufal_medical_corpus

[9] http://www.statmt.org/europarl/

[10] https://globalvoices.org/

News Commentary. The IBECS, Scielo, Pubmed and a part of the MedlinePlus corpus have been obtained from the MeSpEn corpus[11] (Villegas et al. 2018).

In Table 2.1 the size of all corpora and glossaries used for training the MT systems are shown. The figures are calculated after eliminating all the repeated source segment – target segment pairs in the corpora.

Table 2.1: Size of the corpora and glossaries used to create the corpus to train the MT systems.

Corpus	Segments/Entries	Tokens eng	Tokens spa
EMEA	366,769	5,327,963	6,008,543
IBECS	628,798	13,432,096	14,879,220
MedLine Plus	15,689	209,074	234,660
MSD Manuals	241,336	3,719,933	4,467,906
Portal Clinic	8,797	159,717	169,294
PubMed	320,475	2,752,139	3,035,737
UFAL	258,701	3,202,162	3,437,936
Glossary MeSpEn	125,645	286,257	348,415
ICD10-en-es	5,202	25,460	30,580
SnowMedCT Denom.	887,492	3,509,062	4,457,681
SnowMedCT Def.	4,268	177,861	184,574
In-domain	2,836,580	32,479,955	36,893,257
Scielo	741,407	17,464,256	19,305,165
Europarl	1,961,672	50,008,219	52,489,142
Global Voices	559,418	10,717,938	11,496,683
News Commentary	259,412	5,898,912	6,903,975
Out-of-domain	3,521,363	84,087,899	90,193,659

4 Automatic evaluation of the MT systems

In Table 2.2 we can observe the evaluation values of the trained systems using MTEval[12] along with Apertium and Google Translate. This software allows to calculate BLEU, NIST, RIBES and WER using only one reference. We have used all

[11] http://temu.bsc.es/mespen/

[12] https://github.com/odashi/mteval

the test sets of the corpus. As shown in the table, the systems trained in the experiment obtain better results in all metrics than the reference systems used, except for the Google Translate system, which obtains a slightly better NIST result than the MMT Phrase-Based system without context and a better WER result than the two MMT Phrase-Based systems. The MMT Neural system performs consistently better than the MMT Phrase-Based system. In the MMT Neural system, we do not see any significant difference between the results obtained when trained with or without context.

Table 2.2: Results of the automatic evaluation using mteval.

MT system	BLEU	NIST	RIBES	WER
Apertium	0.192577	6.442539	0.713117	0.702716
Google T.	0.402497	9.632268	0.809469	0.530053
MMT P.B. no context	0.424183	9.536248	0.814425	0.637821
MMT P.B. context	0.444832	9.801466	0.819303	0.621032
MMT Neural no context	0.503935	11.106222	0.836954	0.485474
MMT Neural context	0.505778	11.141294	0.836313	0.481039

5 Experiments

We carried out three different experiments with English-Spanish medical texts to assess human perception and evaluation of both PBSMT and NMT systems.

5.1 Translation ranking

In the first part, participants had to answer some questions about their previous experience in the translation industry. The survey was open both to students and professional translators as we were mainly interested in the perception of quality. In the second part of the survey, participants had to rank the translation of 40 segments (human translation, NMT and PBSMT), which had no context and were randomized to avoid bias. They were selected so there were no repeated translations and all had a minimum length of 100 characters. Then we applied a script to ensure there was a minimum editing distance of 15% between the human-PBSMT, human-NMT and PBSMT-NMT solutions. This reduced the number of segments from 230 to 145. We hand-picked 40 segments without typos nor any other problem.

5.2 Fluency and adequacy

We presented a survey with the same English segments as in the previous experiment. In the first part, participants (both students and professional translators) had to answer some questions about their previous experience in the translation industry. Afterwards, they had to evaluate the fluency and adequacy of the proposed translation on a four-point Likert scale. The translation was either PBSMT or NMT chosen randomly without any knowledge of the participants. The goal was to assess fluency and adequacy for in-domain formal texts.

5.3 PE time and technical effort

Finally, in the third experiment, participants had to post-edit 41 segments from a 2018 medical paper. They had to carry out the task in PET (Aziz et al. 2012)[13], a computer-assisted translation tool that supports PE. It was used with its default settings. It logged both PE time and edits (keystrokes, insertions and deletions, that is, technical effort). Four professional translators with more than two years of experience post-editing carried out the task: two of them post-edited the PBSMT output and the other two post-edited the NMT output.

6 Results

6.1 Translation ranking

29 people answered the survey. From those, 86.21% had previous experience as translators and 58.62% had worked on PE tasks. Confirming the initial hypothesis, most respondents preferred the human translation. However, this percentage was only of 60.52%. The second most preferred translation was NMT, with 25.17%, and PBSMT was only considered the best translation for 14.31% of the segments. We calculated inter annotator agreement using Fleiss' kappa (Fleiss 1971), which showed a fair agreement among the annotators ($\kappa = 0.36$). These results were statistically significant in a one-way ANOVA comparison ($p < 0.05$).

Although the survey was conducted on a fairly small number of sentences, it seems to point in two directions: NMT is far from achieving the quality of human translation for medical texts, and NMT yields better translations than PBSMT. We conducted a manual analysis of the sentences in which NMT or PBSMT were selected as the best translation. It was observed the main reason for the selection was terminology precision and fluency of the MT output.

[13] http://wilkeraziz.github.io/dcs-site/pet/index.html

Table 2.3: Results of the human-NMT-PBSMT ranking survey.

Evaluation	Human	NMT	PBSMT
EN-ES (40)	60.52%	25.17%	14.31%

6.2 Fluency and adequacy

In the second experiment, eleven people answered the survey. Seven of them were translators with more than two years of experience and only four of them were students. Both fluency and adequacy obtained a higher rate for NMT after calculating the mean for both MT systems. We calculated inter annotator agreement using Fleiss' kappa (Fleiss 1971). For fluency, it showed poor agreement among the annotators ($\kappa = 0.01$). Results were statistically significant in a one-way ANOVA comparison, with an F-ratio value of 2.75586 and a p-value of 0.04856 (significance at $p < 0.05$). For adequacy, there was also poor agreement among annotators. These results weren't statistically significant, with an F-ratio value of 0.96767 and a p-value of 0.412816 ($p < 0.05$).

If we take a closer look at the sentences that had to be assessed, PBSMT segments often contain morphological problems (e.g. concordance) that we cannot spot in NMT segments, as in example (1). This way the generally higher ratings for fluency and adequacy of the NMT system are confirmed.

(1) Source: Craniopharyngioma had more hormone deficiencies
Gloss: Craneofaringioma tenían más déficits hormonales
PBSMT: 'Craneofaringioma/had (plural)/more/deficits/hormonal'

Table 2.4: Results of the ranking survey.

System	Fluency	Adequacy
PBSMT	2.28	2.24
NMT	2.46	2.50

6.3 PE time and technical effort

Results for the PE task by professional translators have been grouped in temporal effort and technical effort (see Tables 2.5 and 2.6). In both cases, the mean for

PBSMT is higher, though only technical effort shows a statistically significant difference (in a t-test with a p-value of 0.002054). It is worth highlighting that there was a considerable difference in time and keylogging between the translators, especially for the two professionals who post-edited PBSMT (as indicated by the standard deviation in Tables 2.5 and 2.6).

Table 2.5: Temporal PE effort (secs/segment).

System	Mean	SD
PBSMT	88.75	44.59
NMT	79.25	33.43

Table 2.6: Technical effort (keystrokes/segment).

System	Mean	SD
PBSMT	130.68	39.63
NMT	54.99	16.90

7 Conclusions and future work

Although the number of segments analyzed is quite small, for this language combination and text type, there seems to be a clear preference for human translations, which are considered better in more than half of the cases. Regarding MT engines, NMT presents more fluency and adequacy. This corresponds with the higher results in all automatic metrics. However, the results for the perception and automatic assessments do not correlate with PE time, even though there is a reduction in technical effort when post-editing NMT outputs. Thus, even though NMT produces more fluent results, this improvement does not always entail a reduction of the PE effort for professional translators, probably due to the added difficulty of error spotting in more fluent outputs.

In future research, we intend to further analyze PE, increasing the number of segments and language combinations to assess the correlation between automatic metrics and PE (technical and temporal) effort.

Acknowledgements

We want to thank the copyright holders for granting permission for the MSD-Manuals website and for using these texts to create an English-Spanish parallel corpus. The training of the neural MT systems has been possible thanks to the NVIDIA GPU grant programme.

References

Allen, Jeffrey H. 2003. Post-editing. In Harold Sommer (ed.), *Computers and translation: A translator's guide*, 297–317. Amsterdam: John Benjamins. DOI: 10.1075/btl.35.19all.

Aziz, Wilker, Sheila C. M. De Sousa & Lucia Specia. 2012. PET: A tool for post-editing and assessing machine translation. In *Proceedings of the eight international conference on language resources and evaluation (LREC'12)*, 3982–3987.

Bahdanau, Dzmitry, Felix Hill, Jan Leike, Edward Hughes, Pushmeet Kohli & Edward Grefenstette. 2018. Jointly learning "what" and "how" from instructions and goal-states. In *6th international conference on learning representations, ICLR 2018, Vancouver, BC, Canada, April 30–May 3, 2018, workshop track proceedings.* OpenReview.net. https://openreview.net/forum?id=BkmZvdkPM.

Bentivogli, Luisa, Arianna Bisazza, Mauro Cettolo & Marcello Federico. 2016. Neural versus phrase-based machine translation quality: A case study. In *Proceedings of the 2016 conference on empirical methods in Natural Language Processing*, 257–267. Austin, Texas: Association for Computational Linguistics. DOI: 10.18653/v1/D16-1025.

Bojar, Ondřej, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia & Marco Turchi. 2017. Findings of the 2017 conference on machine translation (WMT17). In *Second conference on machine translation*, 169–214. http://www.aclweb.org/anthology/W17-4717.

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 & Marcos Zampieri. 2016. Findings of the 2016 conference on machine translation. *Proceedings of the First Conference on Machine Translation* 2. 131–198. http://www.aclweb.org/anthology/W16-2301.

Bojar, Ondřej, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn & Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the third conference on machine translation*, 272–303. http://aclweb.org/anthology/W18-6401.pdf.

Castilho, Sheila, Joss Moorkens, Federico Gaspari, Iacer Calixto, John Tinsley & Andy Way. 2017. Is neural machine translation the new state of the art? *The Prague Bulletin of Mathematical Linguistics* 108(1). 109–120. DOI: 10.1515/pralin-2017-0013.

Castilho, Sheila, Joss Moorkens, Federico Gaspari, Rico Sennrich, Vilelmini Sosoni, Yota Georgakopoulou, Pintu Lohar, Andy Way, Miceli Barone & Maria Gialama. 2017. A comparative quality evaluation of PBSMT and NMT using professional translators. In *Proceedings of MT summit XVI, vol.1: Research track*, 116–131.

Castilho, Sheila, Sharon O'Brien, Fabio Alves & Morgan O'Brien. 2014. Does post-editing increase usability? A study with Brazilian Portuguese as target language. In *Proceedings of the 17th annual conference of the European association for machine translation*, 183–190. Dubrovnik, Croatia: European Association for Machine Translation. https://www.aclweb.org/anthology/2014.eamt-1.40.

Cho, Kyunghyun, Bart van Merriënboer, Dzmitry Bahdanau & Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8: Eighth workshop on syntax, semantics and structure in statistical translation*, 103–111. Doha, Qatar: Association for Computational Linguistics. DOI: 10.3115/v1/W14-4012.

Fleiss, Joseph L. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin* 76(5). 378–382.

Forcada, Mikel L., Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez & Francis M. Tyers. 2011. Apertium: A free/open-source platform for rule-based machine translation. *Machine translation* 25(2). 127–144.

Germann, Ulrich, Eduard Barbu, M. Bentivoglio, Nikolay Bogoychev, C. Buck, D. Caroselli, L. Carvalho, A. Cattelan, R. Cattoni, Mauro Cettolo, Marcello Federico, Barry Haddow, David Madl, L. Mastrostefano, Prashant Mathur, A. Ruopp, A. Samiotou, V. Sudharshan, M. Trombetti & Jan van der Meer. 2016. Modern MT: A new open-source machine translation platform for the translation industry. *Baltic Journal of Modern Computing* 4. 397–397.

Guerberof, Ana. 2009a. Productivity and quality in MT post-editing. In *Proceedings of MT Summit XII: Beyond translation memories: New tools for translators MT*, 1–9. Ottawa, Canada: AMTA. http://www.mt-archive.info/MTS-2009-Guerberof.pdf.

Guerberof, Ana. 2009b. Productivity and quality in the post-editing of outputs from translation memories and machine translation. *The International Journal of Localisation* 7(1). 11–21. https://www.tdx.cat/bitstream/handle/10803/90247/GuerberofThesis%20Final.pdf?sequence=1&isAllowed=y.

Isabelle, Pierre, Colin Cherry & George F. Foster. 2017. A challenge set approach to evaluating machine translation. *Computing Research Repository* abs/1704.07431. http://arxiv.org/abs/1704.07431.

Junczys-Dowmunt, Marcin, Tomasz Dwojak & Hieu Hoang. 2016. Is neural machine translation ready for deployment? A case study on 30 translation directions. *CoRR* abs/1610.01108. http://arxiv.org/abs/1610.01108.

Koehn, Philipp. 2005. Europarl: A parallel corpus for statistical machine translation. In *Procedings of the Xth MT summit*, vol. 5, 79–86.

Koponen, Maarit. 2012. Comparing human perceptions of post-editing effort with post-editing operations. In *Proceedings of the seventh workshop on statistical machine translation* (WMT '12), 181–190. Montréal, Canada: Association for Computational Linguistics. https://www.aclweb.org/anthology/W12-3123.

Koponen, Maarit. 2016. Is machine translation post-editing worth the effort? A survey of research into post-editing and effort. *The Journal of Specialised Translation* 25. 131–148.

Krings, Hans P. 2001. *Repairing texts: Empirical investigations of machine translation post-editing processes*. Vol. 5. Kent, OH: Kent State University Press.

Lommel, Arle & Donald A. DePalma. 2016. *Europe's leading role in machine translation: How Europe is driving the shift to MT*. Tech. rep. Boston. http://cracker-project.eu.

Moorkens, Joss. 2018. Eye tracking as a measure of cognitive effort for post-editing of machine translation. In Walker Calum & Federico M. Federici (eds.), *Eye tracking and multidisciplinary studies on translation*, 55–70. Amsterdam. DOI: 10.1075/btl.143.04moo.

Moorkens, Joss & Sharon O'Brien. 2013. User attitudes to the post-editing interface. In *Proceedings of machine translation summit XIV: Second workshop on post-editing technology and practice, Nice, France*, 19–25. http://www.mt-archive.info/10/MTS-2013-W2-Moorkens.pdf.

Moorkens, Joss, Sharon O'Brien, Igor A. L. Da Silva, Norma B. De Lima Fonseca & Fabio Alves. 2015. Correlations of perceived post-editing effort with measurements of actual effort. *Machine Translation* 29(3). 267–284. DOI: 10.1007/s10590-015-9175-2.

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

Papineni, Kishore, Salim Roukos, Todd Ward & Wj Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In vol. July, 311–318. DOI: 10.3115/1073083.1073135.

Parra Escartín, Carla & Manuel Arcedillo. 2015. A fuzzier approach to machine translation evaluation: A pilot study on post-editing productivity and automated metrics in commercial settings. In vol. 1, 40–45. https://aclweb.org/anthology/W/W15/W15-4107.pdf.

Plitt, Mirko & François Masselot. 2010. A productivity test of statistical machine translation post-editing in a typical localisation context. *The Prague bulletin of mathematical linguistics* 93. 7–16. https://ufal.mff.cuni.cz/pbml/93/art-plitt-masselot.pdf.

Sanchez-Torron, Marina & Philipp Koehn. 2016. Machine translation quality and post-editor productivity. In *Proceedings of AMTA 2016*, 16–26. https://researchspace.auckland.ac.nz/handle/2292/31486.

Specia, Lucia. 2010. Combining confidence estimation and reference-based metrics for segment-level MT evaluation. In *The ninth conference of the association for machine translation in the Americas*. https://amta2010.amtaweb.org/AMTA/papers/2-03-BanerjeeDuEtal.pdf.

Specia, Lucia. 2011. Exploiting objective annotations for measuring translation post-editing effort. In *Proceedings of the 15th conference of the European association for machine translation*, 73–80. http://www.mt-archive.info/EAMT-2011-Specia.pdf.

Toral, Antonio & Víctor M. Sánchez-Cartagena. 2017. A multifaceted evaluation of neural versus phrase-based machine translation for 9 language directions. In *Proceedings of the 15th conference of the European chapter of the Association for Computational Linguistics*, vol. Volume 1: Long papers, 1063–1073. Valencia, Spain: Association for Computational Linguistics.

Villegas, Marta, Ander Intxaurrondo, Aitor Gonzalez-Agirre, Montserrat Marimon & Martin Krallinger. 2018. The MeSpEN resource for English-Spanish medical machine translation and terminologies: Census of parallel corpora, glossaries and term translations. *Language Resources and Evaluation* 52. 32–39.

Wu, Yonghui, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes & Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR* abs/1609.08144. http://arxiv.org/abs/1609.08144.