Does NMT make a difference when post-editing closely related languages? The case of Spanish-Catalan

Sergi Alvarez
Universitat Pompeu Fabra
salvarezvid@uoc.edu

Antoni Oliver
Universitat Oberta de Catalunya
aoliverg@uoc.edu

Toni Badia
Universitat Pompeu Fabra
toni.badia@upf.edu

Abstract

In the last years, we have witnessed an increase in the use of post-editing of machine translation (PEMT) in the translation industry. It has been included as part of the translation workflow because it increases productivity of translators. Currently, many Language Service Providers offer PEMT as a service.

For many years now, (closely) related languages have been post-edited using rule-based and phrase-based machine translation (MT) systems because they present less challenges due to their morphological and syntactic similarities. Given the recent popularity of neural MT (NMT), this paper analyzes the performance of this approach compared to phrase-based statistical MT (PBSMT) on in-domain and general domain documents. We use standard automatic measures and temporal and technical effort to assess if NMT yields a real improvement when it comes to post-editing the Spanish-Catalan language pair.

1 Introduction

Machine translation (MT) between (closely) related languages presents less challenges and has received less attention than translation between distant languages because it shows a smaller number of translation errors. For a long time now, post-editing of machine translation (PEMT) has been included as a regular practice for these language combinations because it increases productivity and reduces costs (Guerberof, 2009a).

Catalan and Spanish are closely-related languages derived from Latin. They share many morphological, syntactic and semantic similarities. This yields good results for rule-based and statistical-based systems. These systems are currently being used for post-editing both general and in-domain texts in many different companies and official organizations.

The quality of the MT output is one of the main elements that determines the post-editing effort. The higher the MT quality, the more effective post-editing can be. However, automatic metrics generally used to assess the quality of MT do not always correlate to the required post-editing effort (Koponen, 2016). Nor does translators' perception tend to match PE effort (Koponen, 2012; Moorkens et al., 2018). Research in this field has mainly focused on measuring the post-editing effort related to MT output quality (Guerberof, 2009a; Guerberof, 2009b; Specia, 2011; Specia, 2010), productivity (O'Brien, 2011; Parra Escartín and Arcedillo, 2015; Plitt and Masselot, 2010; Sanchez-Torron and Koehn, 2016), translator's usability (Castilho et al., 2014; Moorkens and O'Brien, 2013) and perceived post-editing effort (Moorkens et al., 2015).

Regarding post-editing effort, all research uses the three separated, but inter-related, dimensions established by Krings (2001): temporal, technical and cognitive. Temporal effort measures the time spent post-editing the MT output. Technical effort makes reference to the insertions and deletions applied by the translator and is usually measured with keystroke analysis with HTER (Snover et al., 2006). Cognitive effort relates to the cognitive processes taking place during post-editing and has been measured by eye-tracking or think-aloud protocols. Krings (2001) claimed that post-editing effort could be determined as a combination of all three dimensions. Even though no current measure includes them all, cognitive effort correlates with technical and temporal PE effort (Moorkens et al., 2015).

In recent years, neural MT has gained popularity because the results obtained in terms of quality have been very successful as evidenced in WMT 2016 (Bojar et al., 2016), WMT 2017 (Bojar et al., 2017), and WMT 2018 (Bojar et al., 2018). These results have initiated a shift from statistical machine translation (SMT) to neural machine translation (NMT) in many translation industry scenarios. Google, for example, which first used rule-based MT, and then (phrase-based) SMT, has very recently replaced some of their statistical MT engines by NMT engines (Wu et al., 2016).

As NMT is becoming more popular among language service providers and translators, it is essential to test if it can really improve the post-editing process compared to phrase-based SMT (PSMT). Recent research (Bentivogli et al., 2016; Castilho et al., 2017) has shown an improved quality of NTM for post-editing certain language pairs, such as German, Greek and Portuguese (Castilho et al., 2017). But as far as we know, post-editing closely related languages has been scarcely analyzed before. We carry out two sets of experiments. The first experiments compare the post-editing of NMT and PBSMT output for general news texts from Spanish into Catalan. The second batch of experiments focus on in-domain formal documents and study the post-editing of NMT and PBSMT output for Spanish to Catalan UE documents. The latter texts tend to have more fixed syntactic structures than the former, but present a larger use of technical content and terminology. In both sets of experiments we compare post-editing temporal and technical effort with automatic metrics. We also carry out a manual analysis of the machine translation outputs.

Given the similarities between Spanish and Catalan, we want to test if NMT improves temporal or technical post-editing effort for these two languages. This leads us to the main questions that this paper tries to solve:

- Which MT method (PBSMT or NMT) yields better results for post-editing Spanish into Catalan?
- How do post-editing measures correlate with automatic metrics?
- How does the domain and the formality of the texts affect the post-editing performance between Spanish and Catalan?

2 Related Work

MT systems between related languages have always been considered less complex. In fact, rule-based MT and SMT have yielded better results for these language combinations (Vicic and Kubon, 2015; Kolovratník et al., 2009). In the last few years, there has been an increasing attention on NMT and recent research has tried to analyze if there is a real improvement in quality, both using automatic metrics and human evaluation. Bentivogli et al. (2016) write one of the first research papers comparing how NMT and SMT affect post-editing. They post-edit NMT and SMT outputs of English to German translated TED talks to analyze both results. They conclude that one of the main strengths of NMT is reodering of the target sentence. In general terms, NMT decreases the post-editing effort, but degrades faster than SMT with sentence length.

Wu (2016) compares BLEU (Papineni et al., 2002) and human scores for machine-translated wikipedia entries to evaluate the quality of NMT and SMT. This paper and others (Junczys-Dowmunt et al., 2016; Isabelle et al., 2017) confirm that there is an improvement in the global quality of the translated output using NMT systems.

Toral and Sánchez-Cartagena (2017) take the study by Bentivogli et al. (2016) and increase the initial scope by adding different language combinations and metrics. Although they conclude that NMT produces a better quality than previous systems, the improvement is not always clear for all language combinations.

Castilho et al. (2017) report on a comparative study of PBSMT and NMT. It analyzes four language pairs and different automatic metrics and human evaluation methods. In general, NMT produces better results, although the paper highlights some strengths and weaknesses. It pays special attention to post-editing and uses the PET interface (Aziz et al., 2012) to compare educational domain output from both systems using different metrics. One of the conclusions is that NMT reduces word order errors and improves fluency for certain language pairs, so that fewer segments require post-editing. However, the PE effort is not reduced when working with NMT output.

Koponen et al. (2019) present a comparison of PE changes performed on NMT, RBMT and SMT output for the English-Finnish language combina-

Corpus	Segments	Tokens es	Tokens ca
DOGC	6,943,595	155,233,465	157,000,914
General	4,163,009	93,489,848	93,538,673

Table 1: Size of the training corpora

System	BLEU	NIST	WER
NMT Marian Admin.	0.845	13.055	0.1424
PBSMT Moses Admin.	0.896	13.458	0.0881
Google Translate Admin.	0.869	13.279	0.0918
NMT Marian General	0.767	12.426	0.185
PBSMT Moses General	0.812	12.799	0.171
Google Translate General	0.826	12.980	0,121

Table 2: Automatic evaluation figures

tion. A total of 33 translation students edit in this English-to-Finnish PE experiment. It outlines the strategies participants adopt to post-edit the different outputs, which contributes to the understanding of NMT, RBMT and SMT approaches. It also concludes that PE effort is lower for NMT than SMT.

Regarding NMT for related languages, Costa-Jussà (2017) analyzes automatic metrics and human scores for NMT and SMT from Spanish into Catalan. She concludes that NMT quality results are better both for automatic metrics and human evaluation for in-domain sets, but PBSMT results are better for general domain ones. However, as far as we are concerned, there are no studies analyzing how these MT outputs affect post-editing for in-domain texts, although there have been other papers with a more linguistic approach that have studied the main linguistic issues for NMT between certain related language pairs (Popovic et al., 2016).

3 MT systems and training corpora

For our experiments, we have trained two statistical and two neural machine translation systems: one of each for a general domain and the other for the Administrative/Legislative domain.

3.1 Corpora

For the general domain we have combined three corpora: (1) a self-compiled corpus from Spanish-Catalan bilingual newspapers; (2) the GlobalVoices corpus (Tiedemann, 2012) and (3) the Open Subtitles 2018 corpus (Lison and Tiedemann, 2016).

The systems for the Administrative/Legislative domain have been trained with the corpus from the Official Diary of the Catalan Government (Oliver, 2017). The Catalan part of the corpora has been normalized according to the new orthographic rules of Catalan. This step has been performed in an automatic way.

In Table 1 the sizes of the training corpora are shown. A small part of the corpus (1000 segments) has been reserved for optimization (statistical) and validation (neural). Another set (1000 segments) has been reserved for evaluation. So there are no common segments in the train, validation and evaluation subcorpora.

The corpora have been pre-processed (tokenized, truecased and cleaned) with the standard tools distributed in Moses[1]. The same pre-processed corpora have been used for training the statistical and the neural systems.

3.2 PBSMT system

For the statistical system we have used Moses (Koehn et al., 2007) and trained a system for each of the corpora. We have used a language model of order 5. For the alignment we have used mgiza with grow-diag-final-and.

3.3 NMT system

For the neural machine translation system we have used Marian[2] (Junczys-Dowmunt et al., 2018). We have trained the systems using an RNN-based encoder-decoder model with attention mechanism (s2s), layer normalization, tied embeddings, deep encoders of depth 4, residual connectors and

[1] http://www.statmt.org/moses/

[2] https://marian-nmt.github.io

Domain	System	Mean	Std. Deviation
In-domain (UE)	Marian	50.89	11.78
	Moses	73.70	29.60
	Google	34.68	10.88
General domain	Marian	33.71	2.75
	Moses	42.94	13.96
	Google	32.93	12.65

Table 3: Temporal post-editing effort (secs/segment)

Domain	System	Mean	Std. Deviation
In-domain (UE)	Marian	64.55	65.75
	Moses	12.09	10.50
	Google	2.23	1.38
General domain	Marian	37.99	31.91
	Moses	16.43	1.62
	Google	27.34	37.88

Table 4: Technical post-editing effort (keystrokes/segment)

LSTM cells (following the example of the Marian tutorial[3]).

4 Automatic evaluation of the MT systems

The systems have been automatically evaluated using mteval[4] to obtain the values for BLEU, NIST and WER. Table 2 includes the evaluation figures for all the MT systems used. As a reference, we also include the metrics for Google Translate[5] for the same evaluation sets.

5 Experiments

We have carried two sets of experiments to assess the correlation of MT metrics with the post-editing time and technical effort. The participants were students in their last year of the Degree in Translation and Language Sciences. They post-edited during a PE task organized as part of a course on Localization taught by one of the authors. They all acknowledged a C2 level of both languages. Although students may not be experienced professionals, the participants have translated into this specific language combination during their translation degree program, and have received specific PE training during the course before carrying out the PE task. For these reasons, we can consider them semiprofessionals (Englund Dimitrova, 2005).

In the first experiment, 12 participants post-edited a short text (441 words, 14 segments) from Spanish into Catalan translated with our in-domain PBSMT Moses, our in-domain NMT Marian and NMT Google Translate systems. The text was a passage from a UE document, which presented more fixed syntactic structures, but larger technical content. They had to carry the task using PET (Aziz et al., 2012), a computer-assisted translation tool that supports post-editing. It logs both post-editing time and edits (keystrokes, insertions and deletions, that is, technical effort). As it was a short text, they were asked to post-edit it without any pauses. The main characteristics of the post-editing tool were also explained before beginning the task.

In the second experiment, the same 12 participants post-edited a general domain short text (379 words, 17 segments) from Spanish into Catalan translated with our general purpose PBSMT Moses, our NMT Marian and NMT Google Translate systems. The text was a fragment from a piece of news appeared in the newspaper *El País* on April 4th, 2019. They post-edited the text with the same tool and conditions as in the first experiment.

In order to avoid bias, participants never post-edited the same text twice. We divided the 12 post-editors into groups of 4 people. All the members of each group post-edited the in-domain text translated with an MT system. They also post-edited the general text output for the same MT system.

[3] https://marian-nmt.github.io/examples/mtm2017/complex/

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

[5] Translations were performed on April 9th, 2019

Domain	System	Mean	Std. Deviation
In-domain (UE)	Marian	42.85	0.71
	Moses	53.57	1.50
	Google	85.71	1.32
General domain	Marian	20.59	1.12
	Moses	20.58	1.12
	Google	39.70	0.83

Table 5: Percentage of unmodified segments

6 Results

6.1 Automatic measures

To assess the quality of the MT systems, we included some of the most commonly used automatic evaluation metrics. The BLEU metric (Papineni et al., 2002) and the closely related NIST (Doddington, 2002) are based on n-gram. The word error rate (WER), which is based on the Levenshtein distance (1966), calculates the minimum number of substitutions, deletions and insertions that have to be performed to convert the generated text into the reference text. For all the measurements, our NMT Marian system had the worst rates (see Table 2). However, our PBSMT Moses model had 0.027 BLEU points more than Google Translate for in-domain texts. In the general domain, Google Translate was better rated. That is why we decided to include Google Translate as part of the post-editing tasks.

6.2 Post-editing time and effort

For the in-domain (Administrative/Legislative) post-editing task, our NMT Marian model was the one that took longer post-editing technical effort, although Moses was the one that took longer post-editing temporal effort. This correlates to the worst results in the automatic metrics. In fact, as we can see in the manual evaluation (see example 2, Table 6), errors include adding elements that were not found in the source segment.

Our Moses system had 0.027 BLEU points more than Google Translate in the automatic evaluation. However, post-editors spent less time post-editing the Google Translate output (see Table 3). Regarding the technical effort, Google Translate has a very low rate, which is statistically significant, and correlates to the number of unmodified segments (see Table 5). This correlates to the results obtained by Shterionov et al. (2018), where the automatic quality evaluation scores indicated that the PBSMT engines performed better, but the human reviewers showed the opposite result.

For the general post-editing task, automatic metrics correlate to temporal but not to technical effort. The Google Translate output, which showed a 0.014 increase in BLEU, was translated using far more keystrokes per segment. However, it should be noted the high standard deviation in this case, as in the case of the Marian output.

Another interesting figure is the number of unmodified segments (see Table 5). In this case Google Translate results are far better than Moses, both for in-domain and general domain, which seems to indicate that NMT produces more fluent sentences.

6.3 Manual analysis

The goal of the manual analysis is to complement the information provided by the measures in previous sections. Following Farrús et al. (2010), we have used a taxonomy in which errors are reported according to the different linguistic levels involved: orthographic, morphological, lexical, semantic and syntactic, and according to the specific cases that can be found in the post-editing tasks from Spanish into Catalan. Table 6 shows the error rates for all outputs. Table 7 includes several translation examples from the three systems for the general domain test set. In general, examples show the advantages of the Google Translate neural MT system compared to PBSMT output, in the following terms:

1. There is a **better use of prepositions** in the NMT versions. In this case, the Marian output generates the better version (which includes the pronoun *el* and the use of *el* before the year instead of *en*).
2. There is **a better integrity of meaning** in the Google Translate version. One of the recurrent problems of our Marian version was the addition of extra information or the mistrans-

Domain	System	Ortogr.	Morph.	Lexical	Semantic	Syntactic	Total
In-domain (UE)	Marian	0	0	2	18	0	**20**
	Moses	2	0	2	0	2	**6**
	Google	0	0	0	0	1	**1**
General domain	Marian	0	0	8	5	3	**16**
	Moses	9	12	2	0	5	**28**
	Google	0	11	1	0	3	**15**

Table 6: Number of errors according to the linguistic level

1	ES	Se presume que Van Gogh lo pidió prestado al dueño en 1890 [...]
	Marian	Es presumeix que Van Gogh **el** va demanar prestat al propietari **el** 1890 [...]
	Moses	Es presumeix que Van Gogh **ho** va demanar prestat **el** propietari en 1890 [...]
	Google	Es presumeix que Van Gogh va demanar prestat **a l'amo en** 1890 [...]
2	ES	Es un Lefaucheux [...] hallado en un prado de la localidad de Auvers-sur-Oise por un campesino
	Marian	És **un lladre** [...] trobat en un **enclavament** de la localitat **d'arreu del món**
	Moses	És **un Lefaucheux** [...] trobat en un prat de la localitat **basca** d'Auvers-sud-Oise per un pagès
	Google	És **un Lefaucheux** [...] trobat en un prat de la localitat d'Auvers-sud-Oise per un pagès
3	ES	En 1888, intentaron trabajar juntos en Arlés, al sur de Francia.
	Marian	El 1888, van intentar treballar junts a **Espanya**, al sud de França.
	Moses	En 1888, van intentar treballar junts, a **Arle. Al sud** de França
	Google	En 1888, van intentar treballar junts a Arles, al sud de França.
4	ES	De la pistola no volvió a saberse nada hasta 1965 y su antigüedad está certificada.
	Marian	De la pistola no **es va tornar a saber res** fins al 1965 i la seva antiguitat està certificada.
	Moses	De la pistola no **va tornar a saber res** fins **1965.** Està certificada la seva antigui**tat i**
	Google	De la pistola no **va tornar a saber res** fins a 1965 i la seva antiguitat està certificada.

Table 7: Translation examples

lations, like in this case. The Moses version also ads *basca* (it's the only time Moses adds extra information).

3. The Google Translate version **is more fluent**. Even though the Moses output generally includes all the source information, it sometimes truncates the sentences.
4. NMT achieves **a better syntactic organization** that produces a more understandable sentence with less mistakes.

7 Discussion

This paper shows a comparison between PBSMT and NMT for general and in-domain documents from Spanish into Catalan. Automatic metrics show better results for PBSMT with in-domain texts. However, Google Translate NMT system has a better rate when translating general domain sentences.

Regarding post-editing, for this study, text types, and language pair results show an improvement of unmodified segments and temporal effort for NMT systems. For the in-domain text, with a lower BLUE rate, both technical and temporal effort, as well as the number of unmodified segments and translation errors, show a clear improvement of Google Translate. The manual analysis also confirms that NMT systems tend to solve some of the usual problems of PBSMT systems when translating closely related languages. However, as it is shown in the translation from our NMT Marian system, a lower quality in NMT systems tends to produce unreliable translation outputs, which complicate the post-editing process.

We plan to improve our Marian NMT system using the subword-nmt algorithm (Sennrich et al., 2015) to minimize the effect of out-of-vocabulary words.

Acknowledgments

The training of the neural MT systems has been possible thanks to the NVIDIA GPU grant programme.

References

Aziz, Wilker, Sheila C. M. De Sousa, and Lucia Specia. 2012. PET: A Tool for Post-editing and Assessing Machine Translation. *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 3982–3987.

Bentivogli, Luisa, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. 2016. Neural versus Phrase-Based Machine Translation Quality: a Case Study. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 257–267.

Bojar, Ondrej, 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. 2016. Findings of the 2016 Conference on Machine Translation. *Proceedings of the First Conference on Machine Translation*, 2:131–198.

Bojar, Ondřej, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Lmu Munich, Philipp Koehn, Jhu / Edinburgh, Qun Liu, Varvara Logacheva, Mipt Moscow, Christof Monz, Matteo Negri Fbk, Matt Post, Johns Hopkins, Univ Raphael Rubino, and Marco Turchi Fbk. 2017. Findings of the 2017 Conference on Machine Translation (WMT17). *Proceedings of the Second Conference on Machine Translation*, 2:169–214.

Bojar, Ondřej, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 Conference on Machine Translation (WMT18). *Proceedings of the Third Conference on Machine Translation*, pages 272–303.

Castilho, Sheila, Fabio Alves, Sharon O'Brien, and Morgan O Brien. 2014. Does Post-editing Increase Usability? A Study with Brazilian Portuguese as Target Language. *Proceedings European Association for Machine Translation (EAMT)*, (2010).

Castilho, Sheila, Joss Moorkens, Federico Gaspari, Rico Sennrich, Vilelmini Sosoni, Panayota Georgakopoulou, Pintu Lohar, Andy Way, Antonio Valerio Miceli Barone, and Maria Gialama. 2017. A Comparative Quality Evaluation of PBSMT and NMT using Professional Translators. *Proceedings of MT Summit XVI, vol.1: Research Track*, pages 116–131, 9.

Costa-Jussà, Marta R. 2017. Why Catalan-Spanish Neural Machine Translation? Analysis, Comparison and Combination with Standard Rule and Phrase-based Technologies. *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 55–62.

Doddington, George. 2002. Automatic Evaluation of Machine Translation Quality Using N-gram Co-occurrence Statistics. *Proceedings of the Second International Conference on Human Language Technology Research*, pages 138–145.

Englund Dimitrova, Birgitta. 2005. *Expertise and explicitation in the translation process*. John Benjamins Publishing Company, Amsterdam.

Farrús, Mireia, Marta Costa-jussa, Jose Bernardo Mariño Acebal, and José Fonollosa. 2010. Linguistic-based evaluation criteria to identify statistical machine translation errors. *Proceedings of the 14th Annual Conference of the European Association for Machine Translation*, 01.

Guerberof, Ana. 2009a. Productivity and Quality in MT Post-editing. *Proceedings of MT Summit XII.*

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.

Isabelle, Pierre, Colin Cherry, and George F. Foster. 2017. A Challenge Set Approach to Evaluating Machine Translation. *CoRR*, abs/1704.07431.

Junczys-Dowmunt, Marcin, Tomasz Dwojak, and Hieu Hoang. 2016. Is Neural Machine Translation Ready for Deployment? A Case Study on 30 Translation Directions. *CoRR*, abs/1610.01108.

Junczys-Dowmunt, Marcin, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast Neural Machine Translation in C++. *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, July.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions*, pages 177–180.

Kolovratník, David, Natalia Klyueva, and Ondrej Bojar. 2009. Statistical Machine Translation Between Related and Unrelated Languages. *Proceedings of the Conference on Theory and Practice of Information Technologies, ITAT 2009, Horský hotel Kralova studna, Slovakia, September 25-29, 2009*, pages 31–36.

Koponen, Maarit, Leena Salmi, and Markku Nikulin. 2019. A Product and Process Analysis of Post-editor Corrections on Neural, Statistical and Rule-based Machine Translation Output. *Machine Translation.*

Koponen, Maarit. 2012. Comparing Human Perceptions of Post-editing Effort with Post-editing Operations. *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 181–190.

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*, pages 131–148.

Krings, Hans P. 2001. *Repairing Texts: Empirical Investigations of Machine Translation Post-editing Process*. The Kent State University Press, Kent, OH.

Levenshtein, Vladimir Iosifovich. 1966. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*.

Lison, Pierre and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles.

Moorkens, Joss and Sharon O'Brien. 2013. User Attitudes to the Post-editing Interface. *Proceedings of Machine Translation Summit XIV: Second Workshop on Post-editing Technology and Practice, Nice, France*, pages 19–25.

Moorkens, Joss, Sharon O 'brien, Igor A L Da Silva, Norma B De, Lima Fonseca, Fabio Alves, and Norma B De Lima Fonseca. 2015. Correlations of Perceived Post-editing Effort with Measurements of Actual Effort. *Machine Translation*, 29:267–284.

Moorkens, Joss, Antonio Toral, Sheila Castilho, and Andy Way. 2018. Translators' Perceptions of Literary Post-editing using Statistical and Neural Machine Translation. *Translation Spaces*, 7:240–262.

O'Brien, Sharon. 2011. Towards Predicting Post-editing Productivity. *Machine Translation*, 25(3):197–215.

Oliver, Antoni. 2017. El corpus paral·lel del diari oficial de la generalitat de catalunya: compilació, anàlisi i exemples d'ús. *Zeitschrift für Katalanistik*, 30:269–291.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, July.

Parra Escartín, Carla and Manuel Arcedillo. 2015. A Fuzzier Approach to Machine Translation Evaluation: A Pilot Study on Post-editing Productivity and Automated Metrics in Commercial Settings. *Proceedings of the ACL 2015 Fourth Workshop on Hybrid Approaches to Translation (HyTra)*, 1(2010):40–45.

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

Popovic, Maja, Mihael Arcan, and Filip Klubicka. 2016. Language Related Issues for Machine Translation between Closely Related South Slavic Languages. *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects, VarDial@COLING 2016, Osaka, Japan, December 12, 2016*, pages 43–52.

Sanchez-Torron, Marina and Phillipp Koehn. 2016. Machine Translation Quality and Post-Editor Productivity. *Proceedings of AMTA 2016*, pages 16–26.

Sennrich, Rico, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Shterionov, Dimitar, Riccardo Superbo, Pat Nagle, Laura Casanellas, Tony O'Dowd, and Andy Way. 2018. Human versus automatic quality evaluation of nmt and pbsmt. *Machine Translation*, 32(3):217–235, Sep.

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

Specia, Lucia. 2010. Combining Confidence Estimation and Reference-based Metrics for Segment-level MT Evaluation. *The Ninth Conference of the Association for Machine Translation in the Americas*.

Specia, Lucia. 2011. Exploiting Objective Annotations for Measuring Translation Post-editing Effort. *Proceedings of the European Association for Machine Translation*, (May):73–80.

Tiedemann, Jörg. 2012. Parallel Data, Tools and Interfaces in OPUS. *Lrec*, 2012:2214–2218.

Toral, Antonio and Víctor M. Sánchez-Cartagena. 2017. A Multifaceted Evaluation of Neural versus Phrase-Based Machine Translation for 9 Language Directions. *CoRR*, abs/1701.02901.

Vicic, Jernej and Vladislav Kubon. 2015. A Comparison of MT Methods for Closely Related Languages: A Case Study on Czech - Slovak and Croatian - Slovenian Language Pairs. *Text, Speech, and Dialogue - 18th International Conference, TSD 2015, Pilsen,Czech Republic, September 14-17, 2015, Proceedings*, pages 216–224.

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, and Jeffrey Dean. 2016. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *CoRR*, abs/1609.08144.