Do translator trainees trust machine translation? An experiment on post-editing and revision

Randy Scansani
University of Bologna
Forlì, Italy
randy.scansani@unibo.it

Silvia Bernardini
University of Bologna
Forlì, Italy
silvia.bernardini@unibo.it

Adriano Ferraresi
University of Bologna
Forlì, Italy
adriano.ferraresi@unibo.it

Luisa Bentivogli
Fondazione Bruno Kessler
Trento, Italy
bentivo@fbk.eu

Abstract

Despite the importance of trust in any work environment, this concept has rarely been investigated for MT. The present contribution aims at filling this gap by presenting a post-editing experiment carried out with translator trainees. An institutional academic text was translated from Italian into English. All participants worked on the same target text. Half of them were told that the text was a human translation needing revision, while the other half was told that it was an MT output to be post-edited. Temporal and technical effort were measured based on words per second and HTER. Results were complemented with a manual analysis of a subset of the observations.

1 Introduction

In the last few years, neural machine translation (NMT) has become the state-of-the-art paradigm in the field of machine translation (MT). This fast-paced progress has shaken the translation industry and the research world, causing different reactions. Part of the research world has responded with enthusiastic claims about the quality achieved with this new architecture (Hassan et al., 2018; Wu et al., 2016), while other studies have tempered such enthusiasm, reporting less clear-cut improvements (Toral and Sánchez-Cartagena, 2017; Castilho et al., 2017).

Companies and individual professionals have started to exploit MT more than in previous years. As testified by the 2018 Language Industry Survey[1], for the first time more than half of companies and individual language professionals have stated that they use MT in their workflow. In the same survey repeated in 2019[2], only generic MT engines (Google Translate and DeepL) were chosen among the 20 most-used tools in companies' workflow.

In this uncertain scenario, translators' opinion on MT is likely to be mixed. In the 2019 Language Industry Survey[3], MT was identified as a negative trend by 20% and as a positive one by 30% of the respondents. Lack of training in MT low output quality resulting from adoption of general purpose engines, and a potential downward trend in translation rates may all explain the negative opinion (some) translators have of MT (Läubli and Orrego-Carmona, 2017), and their limited trust, leading to non-adoption of MT suggestions (Cadwell et al., 2018). Investigating how trust towards MT influences translator trainees' behaviour towards the output, along the lines of Martindale and Carpuat (2018), is thus crucial to evaluate the likelihood that translators convincingly embrace MT.

In this contribution, we ask whether translators' trust changes based on the task they are working on, i.e. if they behave differently when they believe they are revising a human translation (HT) *vs.* post-editing an MT output. We see trust as strictly related to productivity: when post-editors/revisers do not trust a text, they are likely to carry out time-consuming and potentially unnecessary searches, or perform unnecessary edits.

In our study, 47 students from a Master's in

[1] A survey on trends in the language industry carried out by EUATC, Elia, FIT Europe, GALA and LINDWeb. https://bit.ly/2RpQtm2

[2] https://bit.ly/2ZknGlL

[3] https://bit.ly/2ZknGlL

translation of an Italian university, revised/post-edited the same English translation of an Italian source text composed of two academic module descriptions. Half of them were told that the translation was an MT output, while the other half was told that the text had been translated by a human translator. We measured the time each participant spent on each sentence, and the number and extent of changes they made. In what follows we summarise previous work on post-editing (PE) and trust (Sect. 2), describe the experimental setting and method (Sect. 3), outline results (Sect. 4) and draw some conclusions (Sect. 5).

2 Related work

2.1 Post-editing of MT

To the best of our knowledge, no work has been published yet on the assessment of trust towards MT as measured in a PE task. Martindale and Carpuat (2018) conducted a survey among non-professionals to understand how their trust was influenced by fluency and adequacy. The former issue is found to have a stronger negative impact on non-professional translators. More recently, Cadwell *et al.* (2018) interviewed two groups of institutional translators to investigate the reasons for adoption or rejection of MT suggestions. Both groups mentioned lack of trust toward MT as one of the reasons for rejecting MT segments.

Focusing on PE tasks in different languages, a number of papers have analysed how performance changes for different subjects or in different work environments, and using one or more effort categories among those listed by Krings (2001): temporal, technical and cognitive. Moorkens and O'Brien (2015) used edit distance and speed to compare the productivity of professionals and students in a PE (En–De) task, whose aim was to evaluate the suitability of the latter for translation user studies. Daems *et al.* (2017) examined how 10 Master's students and 13 professional translators coped with translation from scratch and PE of newspaper articles (En–Nl), measuring translation speed and cognitive load. Moorkens and O'Brien (2015) found that students have a less negative attitude towards technology, but their productivity cannot be compared to that of professionals; by contrast, according to Daems *et al.* (2017) the performance of the two groups was not as different as could be expected, and indeed students were more at ease with PE than professionals.

Yamada (2019) compared perceived cognitive effort, amount of editing and final quality between two PE tasks carried out by students, one using an NMT output and one a PBMT output (En–Ja). While the cognitive effort was similar for the NMT and PBMT tasks, NMT output required less editing effort and led to a better final quality.

Rossetti and Gaspari (2017) measured perceived and real effort of six MA students when translating with translation memories (TMs) and in a PE scenario, triangulating time measurements, think-aloud protocols (TAPs) and retrospective interviews. Results show that only suggestions coming from the TM had a positive impact on perceived task complexity and temporal effort.

Despite growing interest in PE, to the best of our knowledge trust has not been investigated in such task. Furthermore, our language combination (It–En) is relatively under-represented in PE experiments, and the text domain we are focusing on (university module descriptions) is a novel one in this scenario.

2.2 Trust

The notion of trust is a multifaceted one, which has been studied in a host of different fields. McKnight *et al.* (2001) report that, in three different monolingual English dictionaries, on average 17 different definitions of trust are provided. Lee and See (2004) define trust as "the attitude that an agent will help achieve an individual's goal in a situation characterised by uncertainty and vulnerability".

Even though human-machine relationships may develop in the same way as human-human ones (Madhavan and Wiegmann, 2007), the constructs developed to describe trust between human beings do not fully transfer to human-machine interactions (Lee and See, 2004). First, human beings, behave intentionally. Second, interpersonal trust depends on how both parties perceive the counterpart's behaviour, which does not happen when one of the parties involved is a machine. In this case, trust follows from observation of technology performance, from understanding of its underlying architecture, and from intended use (Lee and See, 2004). Translators' lack of trust toward MT might therefore be influenced by different factors, including inconsistency/unpredictability of its output (especially true of NMT), or misconceived expectations about its functioning.

Since several academic programs have recently started to offer courses on MT, the next generation of translators will be the first to enter the market with some knowledge of it. Whether their trust in the technology is likely to increase as a result is still an open question.

3 Experimental setup

3.1 Goals and variables

Post-editors' productivity was analysed with respect to the following variables: *(a)* translation method (students are told that the text is an MT output *vs.* a HT); *(b)* translation correctness (the translation is correct and needs to be confirmed *vs.* it is incorrect and needs to be edited).

3.2 Participants

47 students of the Master's in Specialised Translation of the University of Bologna took part in the experiment. 23 participants worked on the PE task and 24 on the revision task.

Native languages of the participants working on MT were Italian (69.6%), English (4.3%) and other (26.1%). The native language of participants working on the purported revision of a HT was Italian (79.2%), English (8.3%) and other (15.5%). Although translating into English as L2 is not common practice for experiments in this field, the reality of the profession is quite different. Two surveys quoted by Pokorn (2016) revealed, respectively, that for 24% of the respondents the ability of translating into an L2 is essential or important for newly employed translators[4] and that more than 50% of 780 free-lance translators working in 80 states (including Italy) translate into L2 [5]

All students belonged to the same cohort. This allowed us to control for *(i)* their PE/translation experience; *(ii)* their knowledge of the text type and disciplinary domains of the texts; *(iii)* their knowledge of English.

Regarding *(i)*, students attended hands-on modules on CAT tools and on MT and PE as part of their syllabus. One week before the experiment, they received training on the use of MateCat,[6] the tool used for the task (see Sect. 3.3). Also, in a pre-experiment questionnaire, they were asked how much experience they had with the revision of a HT or PEMT in a professional setting. Possible answers were: *None*, *Little*, i.e. from 1 to 5 professional tasks or *Much*, i.e. more than 5 professional tasks. Results are reported in Table 1 and show that the degree of expertise is similar in both groups, since the vast majority of the participants had no or little professional experience. Regarding *(ii)*, all subjects are likely to be familiar with the text type, since course unit descriptions address students, and are unlikely to be acquainted with the domains (pharmacy and chemistry), since their academic background is in languages and linguistics. Concerning *(iii)*, all students are tested upon enrollment in the Master's, a minimum of C1 CEFR being required for admission.[7]

Question	Answers	MT part.	HT part.
Professional experience with MT/PE	None	91.3%	95.8%
	Little	8.7%	0%
	Much	0%	4.2%
MT usefulness for translators	Not useful	0%	0%
	Useful	82.6%	70.83%
	Very useful	17.4%	30.43%

Table 1: Results of the questionnaire on participants' professional experience and opinion on usefulness of MT, split by type of task (HT or PE).

To collect data on participants' opinion regarding MT, in the pre-experiment questionnaire they were asked how useful they thought MT is for translators. Results in Table 1 suggest that all participants have a positive opinion on MT, confirming the results described by Daems *et al.* (2017) and Moorkens and O'Brien (2015) (see Sect 2.1).

3.3 Task

The same text was used for both the MT PE task and the HT revision task. It was composed of two course unit descriptions – for a course on chemistry and one on pharmacy – written in Italian. The English version was produced with a state-of-the-art off-the-shelf NMT system, which ensures the high-quality of the target text used for the experiment.

The final version of the text was the result of a two-step procedure. First, to make sure the text could be believed to be a HT, we checked for possible mistakes typical of MT systems. To establish which sentences were (in)correct, three evaluators were asked to assign each sentence to one

[4] 2011 OPTIMALE survey, involving translation companies from 27 countries – including Italy. `https://bit.ly/2x3V0Bo`

[5] 2014 survey by the International Association of Professional Translators and Interpreters. `https://bit.ly/2h0bjsO`

[6] `https://www.matecat.com/`

[7] `https://bit.ly/2pVyffz`

of the following categories: *(i) correct* (the meaning of the source sentence is conveyed in the target text and no editing is required); *(ii) incorrect* (the meaning of the source sentence is conveyed in the target text but edits are required. In this case, evaluators were asked to annotate the part of the sentence that should be edited); *(iii) wrong* (the meaning of the source sentence is not conveyed in the target text). The final decision as to the correctness of each sentence was made by majority vote. None of the sentences was labelled as *wrong*.

A small amount of edits were performed in order to have half *correct* sentences and half *incorrect* ones in the data set (see Sect. 3.1). At the end of this procedure, the text consisted of 60 sentence pairs, corresponding approximately to 670 source words in total.

Participants worked in MateCat. A project – including a termbase – was assigned to each of them.

A week before, students were given basic information about the experiment.[8] After reading the instructions, students started working autonomously. In the instructions they were invited to work as they normally would. They were asked to deliver a target text of publishable quality, but encouraged to use the provided target text as much as possible and not to over-edit. Researchers were present in the lab throughout.

3.4 Evaluation methods

Productivity was measured in terms of HTER (Snover et al., 2006) between the original text and the participants' edited version, and in terms of words per second (WPS). The latter was obtained by converting MateCat time measurements on a segment level into seconds and dividing them by the number of words in the target text.

Two separate linear mixed models were built, one for each dependent variable, i.e. HTER and WPS. In both cases, the independent variables (or fixed effects) are categorical, i.e. translation method (MT/HT), and translation correctness (correct/incorrect). We included in the model an interaction of the two, with participant and segment as random effects.

Random effects were tested for significance using the likelihood ratio test. Following Gries (2015), a model including all fixed and random effects was built and compared using ANOVAs against different models, each excluding one of the random effects. If the difference between the two models was significant ($p < 0.05$), the random effect was kept in the model.

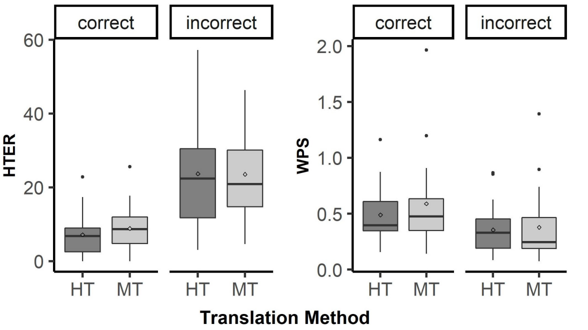

Figure 1: HTER (on the left) and WPS (on the right) values for individual segments split by translation method and correctness of the translation.

4 Results

Tables 2 and 3 summarise significance and estimates for the effects of the two linear mixed models. Figure 1 shows the distribution of HTER and WPS values for individual segments split by translation method and correctness.

4.1 HTER analysis

As expected, in Figure 1 HTER is higher for incorrect sentences overall. While differences between PEMT and HT revision in both cases are small, HTER values for correct MT sentences are slightly higher than values for correct HT sentences.

Moving on to results of our linear mixed model, the two random effects *participant* and *segment* do have a statistically significant impact on the HTER scores (see Table 2), i.e. the observations for the same segment or for the same participant are strongly correlated. Using a mixed model guarantees that the effect of these correlations on the dependent variable is controlled for. Translation correctness is the only fixed effect with a statistically significant impact on HTER, while neither translation method nor its interaction with translation correctness significantly impact on it.

The model thus shows that the number of edits changes significantly only between correct and incorrect sentences, while the amount of edits performed on HT and MT sentences does not differ significantly. The effect of the interaction was not significant either, i.e. no significant change in HTER scores is observed in HT revision and MT PE across translation correctness conditions.

[8] Students were told that the final aim was to compare PE and revision, that data would be collected anonymously and that taking part in the experiment was not compulsory.

		HTER	WPS
	Effect	***p* value**	***p* value**
Random	Participant	<0.001***	<0.001***
	Segment	<0.001***	<0.001***
Fixed	Correctness	<0.0001***	0.1185
	Method	0.6	0.4367
	Interaction	0.14	0.4334

Table 2: Significance of random and fixed effects on the two dependent variables: HTER and WPS

Variables	**HTER**	**WPS**
HT correct	7.284	0.492
MT correct	8.986	0.598
HT incorrect	25.402	0.470
MT incorrect	23.603	0.399

Table 3: Estimates of the two linear mixed models for HTER and WPS. HTER goes up when more edits are performed. WPS goes up when productivity increases.

The similarity of the HTER values is confirmed by estimates in Table 3, where HTER is only slightly higher for MT sentences (+ 1.702), while the opposite happens in incorrect sentences, where HTER is higher for HT revised sentences (+1.799). We conclude that HTER does not provide evidence of a lack of trust toward MT and that behaviours observed for both translation methods are similar.

4.2 Words per second analysis

Figure 1 shows that WPS is higher for correct sentences than for incorrect ones, while it is similar for PE and revision in the two conditions.

As in Sect. 4.1, the *p* values in the WPS column of Table 2 confirm the statistically significant effect of the two random effects (participant and segment) on the dependent variable. However, in this case neither the two fixed effects (translation correctness and translation method), nor their interaction have a significant effect. This means that differences in terms of WPS between correct and incorrect sentences are not statistically significant. Similarly, significant differences between HT revision and PE were not found. When considering the interaction of translation method and translation correctness, WPS does not change significantly.

Looking at Table 3 we can see that, as expected, participants were more productive on correct sentences than on incorrect ones, but values do not vary substantially. WPS is higher (+ 0.106) for correct MT sentences than for correct HT sentences, while for incorrect sentences productivity in terms of WPS is higher (+ 0.010) for HT than for MT.

Combining these results with those in Sect. 4.1, we can confirm that students did not trust MT less than HT or vice versa.

4.3 Qualitative analysis

Given that neither translation method nor its interaction with translation correctness were found to significantly affect technical and temporal effort, we performed a qualitative analysis on a subset of the sentences. Segments with the highest difference between MT and HT in terms of mean HTER were examined.

Concerning Example 1 in Table 4, in both revision and PE, the same number of participants made the right decision, i.e. no edits. In the HT condition most of the participants who edited the sentence only changed the preposition. In the MT condition, terms were changed as well, resulting in a higher HTER score for MT (25.6) than for HT (17.3). Simlarly in Example 2, most post-editors changed verb tenses or nominalised verbs. Mean HTER was 11.4 for MT and 6.79 for HT: most revisers did not edit the sentence.

Regarding incorrect sentences that were edited less in PE than revision, it would seem that revisers paid more attention to issues in the text than post-editors did. For example, all three occurrences of *reaction* in Example 3 should be plural and the term provided by the termbase is *Alkyl halides* rather than *Haloalkane*. 58.3% of the revisers spotted both issues, while only 34.78% of the post-editors did. As a result, mean HTER was 57.2 for HT revision and 43.4 for PE.

In Example 4, it would be sufficient to add the word *examination* at the end. However, in the HT condition most of the participants (54%) carried out a number of other edits applying to the whole sentence. Post-editors carried out unnecessary edits to a lesser extent (4.8%), such that mean HTER was 48.9 for HT and 43.8 for MT.

5 Discussion and limitations

In this contribution we have compared post-editor and reviser trainees' trust towards MT and HT based on HTER and WPS (see Table 2 and 3). According to two linear models, significant changes were only found between HTER on correct and incorrect sentences.

No evidence of a lack of trust towards MT emerged. This behaviour confirms the positive opinion on MT stated in the pre-experiment ques-

Ex.	Sent. type	Text	Correctness
1	OUTPUT	Drugs during pregnancy, in children and in the elderly	Correct
	PE	Drugs in children, in the elderly and during pregnancy	
	REVISION	Drugs during pregnancy, for children and for the elderly	
2	OUTPUT	Finally, possible technical solutions to reduce the use of solvents and their recycling will be discussed	Correct
	PE	Finally, possible technical solutions for solvent usage reduction and solvent recycling will be discussed	
	REVISION	Finally, possible technical solutions to reduce the use of solvents and to enable their recycling will be discussed.	
3	OUTPUT	Haloalkane reactions (metal reaction, elimination reaction)	Incorrect
	PE	Alkyl halides reactions (metal reaction, elimination reaction).	
	REVISION	Alkyl halides reactions (metal reactions, elimination reactions).	
4	OUTPUT	The requirement to take the test is to have taken the Microbiology	Incorrect
	PE	The requirement to take the test is to have taken the Microbiology examination.	
	REVISION	Only the students who passed the Microbiology test can take the exam.	

Table 4: Examples of correct and incorrect outputs with large HTER differences between HT and MT.

tionnaire (see Table 1). This constructive attitude and the ability to interact with technology may be the result of greater awareness of the limits and strengths of MT and PE practice, acquired as part of their academic education (see Sect. 1 and 3.2).

While not significant, differences do exist, and they can provide interesting insights for future work. In correct sentences an increase in HTER corresponds to an increase in WPS – and thus in productivity – and in incorrect sentences a decrease in HTER corresponds to a decrease in WPS. These fluctuations are to be expected, since HTER is based on the number of edits, while WPS is also related to cognitive effort. High HTER scores are often linked to simple preferential changes (see Sect. 4.3), e.g. nominalizations and stylistic vocabulary variation. Such changes may be costly in terms of HTER, but do not require long searches or sentence restructuring – which would be costly in terms of WPS as well. If segments with complex terms are thoroughly checked with a focus on terminology, edits are less costly in terms of HTER than WPS, and discrepancies arise between WPS and HTER. Since participants are not expert in pharmacy or chemistry, terminology searches would not suggest distrust, while preferential changes would. To investigate the presence of preferential changes in the edits, future work might focus on a more thorough qualitative analysis, categorizing the changes introduced in the different conditions and the attention-needing points in the raw output. A longer task would also be necessary, which would however increase fatigue and lead to possible adverse effects, especially since volunteer translator trainees are involved.

In Sect. 3.2 we have seen that students' professional experience is similar in both tasks, and that they are acquainted with the basic notions of PE practice. Their familiarity with revision is certainly greater, though, as this is a standard component in translation courses at both BA and MA level. The more limited familiarity with PE might explain the WPS values obtained, which are highest for MT correct and lowest for MT incorrect. When a mistake is spotted in an MT-translated sentence, more time is spent choosing a strategy to edit it whereas, when a sentence is correct, it is quickly confirmed, as productivity is of the essence in PE. For HT revision, WPS results are more similar in both correctness conditions than is the case in MT. The lowest productivity observed in the MT incorrect condition would suggest that there is still scope for improving translators/post-editors trust in machine translation. More studies would be needed to shed light on the complex and multidimensional nature of trust. For example, pre- and post-experiment questionnaires and interviews could better clarify what participants expect from a HT *vs.* an MT output, and why.

These observations and limitations should not hide the main finding of this study, namely that there are no significant differences between post-editors' and revisers' trust. We would like to interpret this as a sign that, after receiving training on this new technology and before entering the translation industry, a new generation of translators does not seem to be affected by prejudice against PEMT as much as one could expect.

References

Cadwell, Patrick, Sharon OBrien, and Carlos S. C. Teixeira. 2018. Resistance and accommodation:

factors for the (non-) adoption of machine translation among professional translators. *Perspectives*, 26(3):301–321.

Castilho, Sheila, Joss Moorkens, Federico Gaspari, Iacer Calixto, John Tinsley, and Andy Way. 2017. Is neural machine translation the new state of the art? *The Prague Bulletin of Mathematical Linguistics*, 108(1):109–120. Exported from https://app.dimensions.ai on 2018/09/13.

Daems, Joke, Sonia Vandepitte, Robert Hartsuiker, and Lieve Macken. 2017. Translation methods and experience : a comparative analysis of human translation and post-editing with students and professional translators. *META*, 62(2):245–270.

Gries, Th. Stefan. 2015. The most under-used statistical method in corpus linguistics: multi-level (and mixed-effects) models. *Corpora*, 10(1):95–125.

Hassan, Hany, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. Achieving human parity on automatic chinese to english news translation. *CoRR*, abs/1803.05567.

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

Läubli, Samuel and David Orrego-Carmona. 2017. When google translate is better than some human colleagues, those people are no longer colleagues. In *Translating and the Computer 39*, pages 56–69, London.

Lee, John D. and Katrina A. See. 2004. Trust in automation: designing for appropriate reliance. *Human Factors*, 46(1):50–80.

Madhavan, Poornima and Douglas A. Wiegmann. 2007. Similarities and differences between human-human and humanautomation trust: an integrative review. *Theoretical Issues in Ergonomics Science*, 8(4):277–301.

Martindale, Marianna J. and Marine Carpuat. 2018. Fluency over adequacy: A pilot study in measuring user trust in imperfect MT. abs/1802.06041.

McKnight, D. Harrison and Norman L. Chervany. 2001. What trust means in e-commerce customer relationships: An interdisciplinary conceptual typology. *International Journal of Electronic Commerce*, 6(2):35–59, December.

Moorkens, Joss and Sharon O'Brien. 2015. Post-editing evaluations: Trade-offs between novice and professional participants. In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation*, pages 75–81, Antalya, Turkey, May.

Pokorn, Nike K. 2016. Is it so different? competences of teachers and students in 12 translation classes. *International Journal of Translation*, 18:31–48.

Rossetti, Alessandra and Federico Gaspari. 2017. Modelling the analysis of translation memory use and post-editing of raw machine translation output: A pilot study of trainee translators' perceptions of difficulty and time effectiveness. In Hansen-Schirra, Silvia, Oliver Czulo, and Hofmann Sascha, editors, *Empirical Modelling of Translation and Interpreting*, pages 41–67, Berlin. Language Science Press.

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas*, pages 223–231.

Toral, Antonio and 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, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 1: Long Papers*, pages 1063–1073.

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.

Yamada, Masaru. 2019. The impact of Google Neural Machine Translation on post-editing by student translators. *The Journal of Specialised Translation*, pages 87–106, 01.