Machine Translation Markers in Post-Edited Machine Translation Output

Michael Farrell
IULM University
Milan, Italy
michael.farrell@iulm.it

Abstract

The author has conducted an experiment for two consecutive years with postgraduate university students in which half do an *unaided* human translation (HT) and the other half post-edit machine translation output (PEMT). Comparison of the texts produced shows - rather unsurprisingly - that post-editors faced with an acceptable solution tend not to edit it, even when often more than 60% of translators tackling the same text prefer an array of other different solutions. As a consequence, certain turns of phrase, expressions and choices of words occur with greater frequency in PEMT than in HT, making it theoretically possible to design tests to tell them apart. To verify this, the author successfully carried out one such test on a small group of professional translators. This implies that PEMT may lack the variety and inventiveness of HT, and consequently may not actually reach the same standard. It is evident that the additional post-editing effort required to eliminate what are effectively MT markers is likely to nullify a great deal, if not all, of the time and cost-saving advantages of PEMT. However, the author argues that failure to eradicate these markers may eventually lead to lexical impoverishment of the target language.

1 Introduction

To meet the growing demand for translation, the post-editing of machine translation output (PEMT) is being increasingly adopted as a mainstream alternative working method (Koponen, 2016). The compelling reason behind this trend is the widely reported increase in productivity compared to human translation (Aranberri et al. 2014; Plitt and Masselot, 2010) together with a comparable and sometimes higher quality level (Fiederer and O'Brien, 2009; Daems et al., 2017b; O'Curran, 2014; Plitt and Masselot, 2010; Carl et al., 2011). PEMT has been seen to be faster than human translation (HT) for various kinds of text, including non-technical (Daems et al., 2017b), although the increase in productivity in this case is not always statistically significant (Carl et al., 2011).

However, despite the favourable findings regarding PEMT quality, some authors report that readers prefer human translated texts (Fiederer and O'Brien, 2009; Bowker and Buitrago Ciro, 2015). On the other hand, others report that evaluators are not actually able to tell the difference between HT and PEMT (Daems et al., 2017a).

Given the mixed results concerning whether there are any appreciable differences between PEMT and HT, this paper sets out to see if it is possible to identify machine translation (MT) markers in PEMT and therefore design tests to tell them apart.

The primary experiment reported herein was conducted by myself along with 51 postgraduate university students during two consecutive academic years (2016-2017 and 2017-2018) as a classroom exercise designed essentially to reveal:

- The increase in productivity stemming from the use of post-editing.
- The differences between statistical and neural MT output (SMT vs. NMT).
- The existence or otherwise of MT markers in post-edited MT output.

Naturally several other exercises were carried out during the course to analyse other aspects of post-editing and MT, including the building of a custom MT engine (Farrell, 2017).

Proceedings of the 40th Conference Translating and the Computer, pages 50–59,
London, UK, November 15-16, 2018.

I checked all the data the students reported and added several others before analysing them and presenting the results in class. The students involved study the use of machine translation and post-editing at the International University of Languages and Media (IULM) as part of a Master's Degree in Specialist Translation and Conference Interpreting[1].

Besides the much reported increase in productivity, students were expected to find that NMT is better than SMT (Wu et al., 2016), by noting a decrease in post-editing effort (Bentivogli et al., 2016) and therefore time required.

In a comparison between the terminology used in MT, PEMT and HT from English to German, Čulo and Nitzke (2016) observed that HT is more diverse than PEMT in terms of lexical variation, and their results indicated that the MT output *shines through* in PEMT. Students were therefore expected to identify n-grams in the source text which gave rise to a greater variety of translation solutions (TSs) in HT than in PEMT. They were also expected to identify potential MT markers, i.e. TSs which occurred with a statistically significantly higher frequency in PEMT than in HT.

Assuming that they were successful in this, it would then be possible to design tests to distinguish one from the other.

2 Methods

All texts were human translated or machine translated from English into Italian, and the MT outputs were consequently post-edited in Italian. The post-editors were allowed to refer to the source text.

The primary experiment was carried out two years running with groups of postgraduate university students. Approximately half did *unaided* HT and the other half post-edited the MT output obtained from the same texts (total of 51 students). *Unaided* here means the students were not allowed to use translation memory tools, but they could use any dictionaries and web resources they wished.

The experiment was conducted using extracts from the English-language Wikipedia entries describing Venice (153 words) and Verona (168 words), lightly edited to make them consistent as free-standing texts. They were machine translated using Microsoft Translator in November 2016, both in its SMT and NMT versions[2]. The students who translated the text on Venice post-edited one of the two machine translated texts on Verona, and vice versa. They were told to do full post-editing to bring the output up to the same standard as HT, and did not know if they had been given raw SMT or NMT output.

In the first part of the experiment the students measured the time they took for their task.

In the second, they compared their translations with the source text to identify n-grams that had been translated in a wide variety of different ways, and counted the number of ways the same n-gram had been rendered in PEMT. They also checked whether the TS found in the raw MT output was the same as the most commonly chosen TS in HT (top human choice = THC). Moreover they compared the frequency of occurrence of the THCs in the various texts produced.

For reasons explained later, when the raw SMT and NMT outputs proposed the same TS for the n-gram under analysis, the comparison was also made with a combined PEMT group. This is meaningful because the students are faced with essentially the same post-editing choice (leave or change the same raw output TS).

1 Machine Translation and Post-Editing, Course Module Syllabus, International University of Languages and Media (IULM), Milan, Italy: https://bit.ly/2NDrWY2

2 Try & Compare Microsoft's Neural Machine Translation system (no longer available for Italian): https://translator.microsoft.com/neural

The correctness of the TSs chosen was evaluated by ranking them as acceptable, debateable or mistranslations. A mistranslation is a TS declared wrong by agreement. A debateable choice is one which sparked off a potentially endless debate without clear agreement.

Moreover the relative frequency of the THCs was analysed using Fisher's exact two-tailed test. Two by two contingency tables were used (row = THC/all other n-grams chosen; column = HT/PEMT). Debateable choices and mistranslations were omitted from the tables.

The same texts and raw MT outputs were used each year, but the tasks were carried out using different tools. During the first year, the students used Microsoft Word and timed themselves by taking note of the start and finishing times. They also used Microsoft Word tables to compare the various texts, identify n-grams, and write notes. This proved to be a clumsy way of completing the experiment, which spurred me to design a simple software tool, called Raw Output Evaluator (ROE), for the second year (Farrell, 2018). ROE splits the text into segments and displays it in a similar way to a typical Translation Environment Tool, but without the other common CAT tool/TM system functions. Moreover, unlike classic CAT tools, it includes a built-in task timer. It was also used by the post-editors as a simple post-editing interface.

In preparation for this paper, I conducted two additional experiments using the n-grams identified during the course module. In the first of these, I put together texts containing 20 occurrences of the same n-gram using blocks of sentences taken from Wikipedia, and fed them into different free online MT engines (Google Translate[3] and Microsoft Translator[4] in June 2018, and DeepL[5] in August 2018) to get a measure of the variety of different solutions produced in raw MT output for the chosen n-grams. Wikipedia was chosen again for consistency with the primary experiment. The Wikipedia entries where selected using Google (*n-gram site:wikipedia.org*). Blocks normally consisted of whole paragraphs, sometimes shortened a little. Since even neural MT systems seem to choose one of the most statistically frequent HT solutions repeatedly, I expected variety to be low and the THC to occur with a very high frequency.

In the second, I designed a test using a 273-word text extracted from the Wikipedia entry on Venice (lightly edited to make it consistent as a free-standing text) containing five occurrences of the source language translation of a candidate MT marker. I then recruited six professional translators through the Internet (Langit[6] and It-En[7]) and split them into two groups strictly in the order in which they volunteered. One group provided a HT and the other post-edited the Google-translation of the same text (June 2018). The volunteers were told their work was for publication, and that they should therefore aim for an appropriate quality level.

I expected the THC identified by the students to be the most frequently occurring solution in the raw MT output, and this TS to occur with a much greater frequency in the post-edited texts. If the test worked, I expected the three texts with lowest THC frequency to be the HT ones, and the three with the highest frequency to be the post-edited ones. I did not know what degree of variety to expect among the translators but, since the goal of post-editing is to get the job done faster and not waste time making unnecessary edits, I expected any lexical variety observed to be in the translations rather than in the PEMT outputs.

3 https://translate.google.com
4 www.bing.com/Translator
5 www.deepl.com/translator
6 www.turner.it/T-Langit.htm
7 https://groups.yahoo.com/neo/groups/it-en/info

3 Results

3.1 Primary Experiment – HT Time vs. PEMT Time

Tables 1 and 2 only show the results for the first academic year since a bug in the timer function of the software tool used (now fixed) made the second year data unreliable.

Task	Students	Mean time (minutes)	Standard Deviation	Productivity increase
Human Translation	14	19.07	± 5.06	-
Post-editing of SMT	7	18.43	± 7.28	3.47%
Post-editing of NMT	6	18.00	± 9.14	5.94%

Table 1: Time taken to translate or post-edit the Venice text

Task	Students	Mean time (minutes)	Standard Deviation	Productivity increase
Human Translation	13	20.69	± 4.68	-
Post-editing of SMT	7	19.00	± 8.43	8.89%
Post-editing of NMT	7	18.00	± 4.32	14.94%

Table 2: Time taken to translate or post-edit the Verona text

PEMT was faster on average than HT in every case and the post-editing of NMT was faster on average than that of SMT. However the small differences suggest no clear advantage of either MT technology, and the productivity gains are not particularly high. This may depend on the kind of text chosen (see also Carl et al. 2011).

3.2 Primary Experiment – MT Markers

For reasons of time and abundance of data, only the Venice text was analysed for MT markers. To maximize the reliability of the results, the data from both years were put together (total of 50 students – one HT was left out due to an oversight).

The students and I identified 41 n-grams which were judged by rapid observation to have been translated in a greater variety of ways than in the PEMT texts.

There were 26 students in the HT group, 12 in the SMTPE group and 12 in the NMTPE group (a total of 24 students in the combined PEMT group). The first analysis consisted of simply counting the number of different correct TSs used for each n-gram in each group, excluding translation errors. The HT group was compared to the combined PEMT group to have more evenly sized samples (only 25 n-grams were translated in the same way in both raw MT outputs). This comparison was not made between HT and the non-combined PEMT groups because the number of TSs per student (NTS/S) is artificially higher in smaller groups. This is explained by noting that the maximum value of the NTS/S is always one (each student chooses a different solution), but the minimum value (all students choose the same solution) is inversely proportional to the number of students, thus making the smaller group look artificially more inventive than the larger one as we approach the minimum. In more mathematical terms, the assumption that the relationship between number of TSs and group size is linear is false, but it may be a useful approximation when the groups are more or less the same size, hence the need to put the two PEMT groups together.

Of the 25 n-grams therefore considered, the NTS/S was higher in the HT group in 22 cases (88%) and higher in the PEMT group in only 3 (*luxury*, *the fact that*, and *the most notable*). Of the latter three cases, only *luxury* looks significant (2 HT solutions vs. 4 PEMT solutions). The second is virtually a tie (4 solutions/26 students vs. 4 solutions/24 students), and the third is caused by 5 PEMT solutions being disqualified as mistranslations, thus reducing the PEMT

group from 24 to 19 students. The highly uneven group sizes in this case may have distorted the result.

In the 22 cases with greater variety of solutions in the HT group, the NTS/S was more than five times greater in one case (*However*), more than quadruple in another 2 cases (*numerous attractions* and *mainly*), more than triple in another case (*destination*) and more than double in another 4 (*there are*, *people*, *several problems* and *by some*). This therefore confirms our expectation of a much greater variety of TSs in the HT group than in the combined PEMT group.

Moreover we also checked to see if the TS found in the raw MT output was the THC. This was true in 14 cases (56%) in the combined PEMT group. In the other 11 cases, three were the second to top human choice (STHC), one was a different inflection of the THC, two were mistranslations, and one was a solution which all except one of the post-editors chose to change, although strictly not a mistranslation (an unappealing solution). The other 4 were correct solutions that did not rate among the top human choices (16%). Analysis of the 16 cases where the two raw MT outputs contained different TSs revealed that the top plus second to top human choices predominate. In brief, the raw MT outputs more often than not propose the most commonly chosen TSs found in HT.

Fisher's exact two-tailed test was then carried out to see if there were significant differences in the frequency of the THC in the texts produced. This test is able to compensate to some extent for unevenness in group sizes. Considering the combined PEMT group first, in all 9 cases (9/14 = 64%) where the use of the THC was statistically significantly higher in PEMT, the raw MT output contained the THC, which is hardly surprising. In the 5 cases where the use of the THC was statistically significantly lower in PEMT, the raw output contained a mistranslation in one case, the STHC in two, the joint STHC in one (*numerous inhabitants*) and a not particularly high rated alternative solution in only one case. The lower use of the THC is clearly due to the proposal of a highly valid alternative (STHC), except in two cases. Turning to the remaining n-grams and starting with the SMTPE group, there were 2 cases where the use of the THC was statistically significantly higher: the raw output contained the THC in one and a mistranslation in the other. It is not clear why correcting a mistranslation should lead to using the THC more often than usual, also because the opposite was seen in one case in the combined PEMT group. In the SMTPE group there were also 4 cases where the use of the THC was statistically significantly lower. They were all cases where the raw output contained the STHC, which can be explained as before. Concluding with the NMTPE group, in all 3 cases where the use of the THC was statistically significantly higher, the raw output contained the THC. In the only case where the use of the THC was statistically significantly lower (*has caused*), the raw output contained the joint STHC.

In short, there are two predominant cases when there was a statistically significant difference in the frequency of the THC: when the raw MT output contained the THC, in which case it was higher, and when the raw output contained the second to top human choice (STHC), in which case it was lower. This is perfectly in line with expectations and the principle that if a post-editor finds a highly appealing TS (THC or STHC), they tend to leave it and not waste time looking for alternatives.

N-gram	**Raw MT output**	**Statistically significant difference in frequency of THC**			**Greater NTS/S (x greater)**	**Frequency of THC in HT (%)**
		SMT group	**NMT group**	**Combined MT group**		
Today	THC	Very>	Not quite>	Very>	HT	42.31
there are	THC		Extremely>	Very>	HT (x2)	38.46
numerous attractions	THC	Very>	Very>	Extremely>	HT (x4)	34.62

such as	THC	Very>	Yes>	Extremely>	HT	38.46
popular	JTHC/-			n/a		24.00
luxury	THC				PEMT	86.96
destination	THC		Yes>	Yes>	HT	50.00
attracting	BT	Not quite<		Yes<	HT (x 3)	38.46
thousands	THC				HT	76.92
mainly	STHC	Yes<	Yes<	Extremely<	HT (x4)	41.18
people	THC	Not quite>	Extremely>	Very>	HT (x2)	26.92
movie industry	-/THC	Not quite<	Extremely>	n/a		44.00
relies	-/BT			n/a		28.57
heavily	STHC/-	Yes<		n/a		65.22
cruise business	(*)/BT			n/a		25.00
Cruise Venice Committee	BT/BT			n/a		100.00
has estimated	THC	Not quite>			HT	73.08
cruise ship passengers	DI/BT			n/a		52.17
annually	STHC/-			n/a		42.31
in the city	STHC	Very<		Yes<	HT	48.00
However	THC			Yes>	HT (x5)	76.92
major	-/THC			n/a		22.73
worldwide	-		Yes<	Yes<	HT	30.77
tourist destination	-			Not quite<	HT	23.08
has caused	THC/-		Very<	n/a		76.92
several problems	THC	Not quite>	Very>	Very>	HT (x2)	56.00
including	THC/-	Yes>		n/a		52.00
the fact that	THC				PEMT	65.38
very overcrowded	-				HT	23.08
at some points of the year	(**)				HT	48.00
is regarded	DI				HT	42.31
by some	THC	Not quite>		Yes>	HT (x2)	48.00
tourist trap	THC/STHC	Not quite>		n/a		70.83
competition	STHC/THC	Yes<		n/a		46.15
foreigners	THC				HT	84.62
has made prices rise	BT/JTHC	Extremely>	Yes>	n/a		11.54
numerous inhabitants	-	Yes<		Yes<	HT	37.50
to move	STHC/THC	Extremely<	Not quite>	n/a		73.08
more affordable	STHC		Not quite<	Not quite<	HT	26.92
areas	STHC/THC	Extremely<	Yes>	n/a		65.38
the most notable	BT				PEMT	15.38

*Although not strictly a mistranslation, all post-editors chose to change it.
**Although not strictly a mistranslation, all but one post-editor chose to change it.
DI=Different inflection of THC
JTHC = Joint top human choice
STHC = Second to top human choice
BT = Mistranslation (bad translation)

Table 3: Analysis of the 41 n-grams identified

It was decided that an MT marker which might be used to design a test able to distinguish HT from PEMT was one where:

- The THC was found in both kinds of raw MT output
- The THC occurred a very or extremely statistically significant number of times more in PEMT, and
- There was a two or more times greater NTS/S in HT, so it was likely that a greater variety of solutions would also be seen in the test HT.

Four n-grams met these conditions (*there are*, *numerous attractions*, *people* and *several problems*).

3.3 Translation Errors

Errors were only counted for the n-grams analysed, which however amounted to a large proportion of the text (75/153 words = 49%).

	HT	**PEMT**
Debatable choices	18	12
Mistranslation	35	42
Total	53	54
Errors per translator	2.04	2.25

Table 4: Errors found in texts

The PEMT texts were taken together regardless of what the TSs in the raw MT outputs were. The difference between the two groups is not statistically significant whether we count the two kinds of error as separate categories (chi-squared: p=0.35) or lump them together (Fisher's exact two-tailed test: p=0.62). This substantially confirms our expectation that the quality of the two kinds of work is comparable if we evaluate it purely in terms of translation errors.

3.4 First Additional Experiment

The texts analysed contained 20 occurrences each of three of the four MT markers considered ideal for use in the second additional experiment. *People* was excluded because virtually all the top Google hits from Wikipedia used the word in its highly specific meaning of ethnic group or nation (pl. peoples), rather than as the plural of the word person.

N-gram	**Most frequent translation found in raw MT output**	**Microsoft Translator**	**Google Translate**	**DeepL**
There are	Ci sono	20/20 (100%)	18/20 (90%)	18/20 (90%)
Numerous attractions	Numerose attrazioni	19/20 (95%)*	20/20 (100%)	20/20 (100%)
Several problems	Diversi problemi	17/20 (85%)	15/20 (75%)*	18/20 (90%)*

* One of the solutions was a mistranslation

Table 5: Variety of solutions found in raw MT output

Google Translate provided three correct alternatives for *several problems*. In all other cases, only one correct alternative was found. As expected, the variety of TSs for the n-grams studied was low.

The frequency of the THC was extremely statistically significantly higher than in the HTs produced in the primary experiment in the cases of *there are* and *numerous attractions*. In the case of *several problems*, the difference was only very statistically significant in the case of DeepL, not quite statistically significant for Microsoft Translator and not statistically significant for Google Translate. *There are* and *numerous attractions* are therefore the best candidate MT markers for the second additional experiment. *There are* was chosen for its ubiquity, which makes it easily repeatable in a relatively short text without it seeming artificial.

Interestingly, although DeepL is reported by some to give better quality raw MT output than Google Translate (Isabelle and Kuhn, 2018), it would seem to suffer from the same lack of TS variety as the others, if not more so.

3.5 Second Additional Experiment

A 273-word text containing five occurrences of *there are* was given to three professional translators for translation, and Google-translated and given to another three for full post-editing. As was predictable, the raw MT output contained the same TS (*ci sono*) for each occurrence.

	Professional experience (years)	**Time (minutes)**	**Number of occurrences of ci sono**	**Number of different solutions chosen**	**HT/PEMT**
SC	8	51	0	5	HT
LZ	11	32	0	4	HT
MLD	25	64	0	3	HT
CP	16	47	1	5	PEMT
PV	28	45	1	4	PEMT
DG	26	16	4	2	PEMT

Table 6: Results of the *there are* test

The average time taken was 49.00 ± 16.09 minutes for translation and 36.00 ± 17.35 minutes for post-editing, again confirming expectations. None of the volunteers who did the HT translated *there are* with *ci sono*, whereas all the post-editors left at least one occurrence of *ci sono*. Therefore, on this occasion, the test was 100% accurate in distinguishing PEMT from HT. Surprisingly, despite this result, the variety of different TSs chosen in the two groups seems to be comparable, contrary to expectations.

4 Discussion

The primary experiment was not designed solely to identify MT markers. Consequently, result analysis proved quite complex, particularly due to the uneven group sizes.

However the results confirm what would be expected from simple reasoning:

- When a post-editor is faced with an acceptable solution in raw MT output they tend to leave it unedited, even if it is only one of many possible valid solutions.
- Due to the way it works, MT tends to choose one of the solutions most frequently chosen by translators (THC or JTHC).
- Therefore the statistically most frequent solutions in HT occur with a higher than natural frequency in PEMT (MT markers).
- MT markers may be used to design tests to distinguish HT from PEMT.

This experiment also says nothing about the range of solutions used by a single translator or post-editor for a repeated n-gram, but rather the variety chosen by a group of translators or post-editors. It would seem reasonable to assume that freedom from a suggested MT solution would allow translators to give rein to a wide variety of solutions, and this is in line with the result for the translators in the second additional experiment (the *there are* test). However, despite the evident influence of the proposed MT solution, the post-editors in the test appear to have come up with a comparably wide range of solutions. This seems rather hard to explain since it means that they deliberately altered several correct n-grams, contrary to the aims of post-editing. In this case however, a different factor may have come into play. Italians are taught that good writers should avoid unnecessary lexical repetition. Five occurrences of the same expression in four paragraphs may have triggered a *repetitiveness alarm*, turning an otherwise correct solution into an unacceptable one. Alternatively it may also be more simply argued that the scale of the second additional experiment may not be big enough to give reliable results.

It would therefore be advisable to repeat the experiment on a larger cohort using much longer texts with more numerous and sparsely repeated MT markers.

Variety and inventiveness are not always desirable features in every kind of text. For example, excessive lexical variation might make a smartphone user's guide more difficult to follow. Nevertheless, there are various other kinds where lexical uniformity would make the text less interesting to read and less intellectually stimulating (marketing, advertising, literature, journalism, education, entertainment, and creative writing in general). In these cases, counting errors and measuring fluency and adequacy are not sufficient to judge translation quality.

What the findings of the primary experiment show however is an apparent normalization and homogenization of the choices made by post-editors as a whole. This may explain why some authors report that HT is judged to be better in terms of style (Fiederer and O'Brien, 2009). One solution might be to program NMT engines to sometimes randomly pick the second or third best fit translated sentence vectors.

Failure to remedy this homogenization may eventually lead to lexical impoverishment of the target language, particularly in cultures where English has become the primary working language in which new written material is created. Obviously it would be possible to train post-editors to add originality and inventiveness to their work by purposely editing parts where there are no formal errors, but this clearly defeats the object of post-editing.

5 Conclusions

There is clear evidence of a homogenization and normalization phenomenon in connection with post-editing. There is also evidence of a decrease in the variety of different solutions chosen, when considering post-editors together as a group, although it was not possible to confirm this when observing the behaviour of post-editors individually.

As MT systems improve - if this means get better at homing in on the most frequently occurring expressions - the homogenization effect will probably be aggravated.

On account of the findings reported herein, the use of PEMT for texts where variety, originality and inventiveness are quality factors would appear to be unadvisable with the MT technology currently available.

Acknowledgements

All trademarks and trade names are the property of their respective owners.

References

Aranberri, Nora, Gorka Labaka, Arantza Diaz de Ilarraza, et al. (2014): Comparison of Post-editing Productivity Between Professional Translators and Lay Users. Paper presented at the AMTA 2014 3rd Workshop on Post-editing Technology and Practice (WPTP-3), Vancouver, Canada, October 26, 2014.

Bentivogli, Luisa, Arianna Bisazza, Mauro Cettolo and Marcello Federico (2016): Neural versus Phrase-Based Machine Translation Quality: a Case Study. ArXiv e-prints.

Bowker, Lynne and Jairo Buitrago Ciro (2015): Investigating the usefulness of machine translation for newcomers at the public library. Translation and Interpreting Studies. 10(2):165-186.

Carl, Michael, Barbara Dragsted, Jakob Elming, et al. (2011): The process of post-editing: A pilot study. Proceedings of the 8th International NLPCS Workshop: Samfundsliterattur, 131-142.

Čulo, Oliver, Jean Nitzke (2016): Patterns of Terminological Variation in Post-editing and of Cognate Use in Machine Translation in Contrast to Human Translation. Baltic J. Modern Computing, Vol. 4 (2016), No. 2, 106-114.

Daems, Joke, Orphée De Clercq and Lieve Macken (2017a). Translationese and post-editese: How comparable is comparable quality? Linguistica Antverpiensia, New Series: Themes in Translation Studies, 16, 89–103.

Daems, Joke, Sonia Vandepitte, Robert J. Hartsuiker, Lieve Macken (2017b): Translation Methods and Experience: A Comparative Analysis of Human Translation and Post-editing with Students and Professional Translators. Meta: Journal des traducteurs/Meta: Translators' Journal.

Farrell, Michael (2017): Building a Custom Machine Translation Engine as part of a Postgraduate University Course: a Case Study. Proceedings of the 39th Conference Translating and the Computer, pages 35–39,

London, UK, November 16-17, 2017.

Farrell, Michael (2018): Raw Output Evaluator, a Freeware Tool for Manually Assessing Raw Outputs from Different Machine Translation Engines. Paper to be presented as a non-commercial workshop at the Translating and the Computer 40 Conference, London, United Kingdom, 15-16 November, 2018.

Fiederer, Rebecca and Sharon O'Brien (2009). Quality and Machine Translation: A realistic objective? The Journal of Specialised Translation, 11, 52–74.

Isabelle, Pierre and Roland Kuhn (2018): A Challenge Set for French --> English Machine Translation. ArXiv e-prints.

Koponen, Maarit (2016): Is machine translation post-editing worth the effort? A survey of research into post-editing and effort. JoSTrans 25, 131–148.

O'Curran, Elaine (2014): Translation quality in post-edited versus human-translated segments: A case study. Paper presented at the AMTA 2014 3rd Workshop on Post-editing Technology and Practice (WPTP-3), Vancouver, Canada, October 26, 2014.

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. 93:7-16.

Wu, Yonghui, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi (2016): Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. ArXiv e-prints..