Time is Everything: A Comparative Study of Human Evaluation of SMT vs. NMT

Emmanuelle Esperança-Rodier, Caroline Rossi

Time is Everything: A Comparative Study of Human Evaluation of SMT vs. NMT

Emmanuelle Esperança-Rodier
Univ. Grenoble Alpes, CNRS,
Grenoble INP*, LIG, 38000 Grenoble,
France
Emmanuelle.Esperanca-Rodier@univ-grenoble-alpes.fr

Caroline Rossi
Univ. Grenoble Alpes, ILCEA4,
38000 Grenoble,
France
Caroline.Rossi@univ-grenoble-alpes.fr

Abstract

Translation process research has developed tools to gather and analyse empirical data, but while a variety of measures have proved useful and reliable to assess machine translation post-editing effort (see e.g. Vieira 2016: 42), translation processes are seldom considered when assessing the relevance of a given Machine translation post-editing (MTPE) scenario.

Our study seeks to determine the impact of including MTPE in the evaluation process. We selected adequacy and fluency ratings. Based on two distinct experimental conditions, we then compared the ratings produced without performing PE and those produced immediately after a light PE process.

Inter-rater reliability was assessed for each segment in each text (N=55) using Fleiss' kappa for adequacy and fluency scores, and an intra class correlation coefficient (Vieira 2016: 52) for temporal measures. While the reliability of the measures collected without PE was low, the measures collected in PET were for the most part homogeneous. Qualitative analyses of the problematic segments, as evidenced by both kappa and intra class correlation coefficients, showed strong Spearman's correlations, whether positive or negative, between temporal measures and all the other metrics for NMT but weakest ones for SMT. Based on these results, we discuss the advantages and risks of NMTPE.

1 Introduction

Machine translation evaluation (MTE) is performed differently and with different goals in academia and industry (Drugan 2013, in Castilho et al. 2018: 11). However, with the current integration of neural machine translation into human translation workflows, reliable measures of the amount of effort needed to post-edit machine translation (PEMT) outputs have become a common goal for researchers, language service providers and machine translation vendors (ibid., p. 29). Translation process research has developed tools to gather and analyse empirical data, but while a variety of measures have proved useful and reliable to measure PEMT effort (see e.g. Vieira 2016: 42), translation processes are seldom considered when assessing the relevance of a given MTPE scenario.

Against this background, our study seeks to determine the impact of including MTPE in the evaluation process. We selected two of the most commonly used scales for the "declarative evaluation" of MT (Humphreys et al. 1991, in Way 2018b: 164): adequacy and fluency ratings. Based on two distinct experimental conditions, we then compared the ratings produced without performing PE and those produced immediately after a light PE process.

* Institute of Engineering Univ. Grenoble Alpes

2 Methodology

Data was collected with a group of 14 trainee translators, using two different text types and two different tools. Based on the requirements of our French curriculum in specialised translation, we selected only translation into French, and for the sake of comparison we kept only one source language: English. A first series of assessments was conducted with KantanMT's language quality review system (LQR), which allows for a simple comparative evaluation of two systems without post-editing the outputs. The second series was done a few weeks later, in Post-Editing Tool (PET, Aziz et al. 2012). Each experimental condition includes two source texts from two different domains (environmental discourse and patents). We generated usable SMT and NMT outputs using the European Commission's MT (eTranslation) with environmental texts and WIPO translate with patent extracts. In both conditions, the students were given a realistic scenario -- i.e. they performed the evaluation, with a view to determining whether the MT output was relevant to a particular order.

2.1 Data

We have selected four documents dealing with two different subjects. Two documents were Patent extracts while the two other two were environmental texts.

Each document has been translated with Machine Translation (MT) systems trained on the domain, thus using WIPO Translate and eTranslation respectively for Patents and environmental texts. As those two MT systems offer a Neural version (NMT) as well as a Statistical one (SMT), we used them both to translate our documents.

Consequently, we obtained two translation versions of the same source document, which are one WIPO SMT translation version and one WIPO NMT version for a patent document. We therefore got four translations for the two Patent extracts: two SMT translations and two NMT for each text.

In the same way, we performed both NMT and SMT translations on the two environmental texts, having thus one eTranslation SMT translation version (called "LegacyMT@EC" in the eTranslation portal) and one eTranslation NMT version (called "cutting edge") for each environmental document, again resulting in four translations.

As a result, we have got 8 documents, as shown in Table 1 below.

Documents	**WIPO NMT**	**WIPO SMT**	**eTranslate NMT**	**eTranslate SMT**	**Total**
Patent 1	1	1			2
Patent 2	1	1			2
Environmental discourse Text 1			1	1	2
Environmental discourse Text 2			1	1	2
Total	2	2	2	2	8

Table 1: data.

Those documents were used in the two assessments conducted with our translator trainees. The first assessments have been conducted using the Kantan tool in order to assess adequacy and fluency of one patent NMT and one SMT translation and of one environmental text based on both NMT and SMT translations.

The second assessment has been fulfilled using PET in order to post-edit translations before assessing adequacy and fluency. For this last assessment, as required in PET (Aziz et al., 2012), we have had to mix the NMT and SMT translations in the same document so that the translator trainees did not know which translation was the NMT one, nor the SMT one. We have indeed, created from the NMT version translation and the SMT version translation, two mixed-version

translation documents as stated in Table 2 hereafter. To do so, we mixed the odd sentences from the NMT version with the even sentences from the SMT version and conversely the even sentences from the NMT version with the odd sentences from the SMT version, to create the second document.

Patent 1 WIPO NMT	**Patent 1 WIPO SMT**	**Mixed Text 1**	**Mixed Text 2**
a-L'invention concerne un ensemble presse-frein (1) comprenant :	A-L'invention concerne un agencement de presse de frein (1) comprenant :	a-L'invention concerne un ensemble presse-frein (1) comprenant :	A-L'invention concerne un agencement de presse de frein (1) comprenant :
b-L'invention concerne également une presse plieuse (2) destinée à plier des pièces, en particulier une presse à border,	B-Une presse de frein et de pliage de pièces, en particulier une presse plieuse,	B-Une presse de frein et de pliage de pièces, en particulier une presse plieuse,	b-L'invention concerne également une presse plieuse (2) destinée à plier des pièces, en particulier une presse à border,
c-un siège (3) positionné devant la presse de frein (2) pour un opérateur de ladite presse de frein (2),	C-Un siège (3) positionnée devant la presse et frein pour un opérateur dudit frein, presse (2)	c-un siège (3) positionné devant la presse de frein (2) pour un opérateur de ladite presse de frein (2),	C-Un siège (3) positionnée devant la presse et frein pour un opérateur dudit frein, presse (2)
d-Et-un système de support réglable (10) qui supporte le siège (3) et qui est fixé à un point de fixation (9), la presse de frein (2) ayant un cadre (4) qui supporte un premier porte-outil (5) et un second porte-outil (6).	D-Et un système de support réglable (10) qui supporte le siège (3) et est fixée au niveau d'un point de fixation (9) (2), le frein de presse ayant un cadre (4) qui supporte un premier porte-outil (5) et un second porte-outil (6).	D-Et un système de support réglable (10) qui supporte le siège (3) et est fixée au niveau d'un point de fixation (9) (2), le frein de presse ayant un cadre (4) qui supporte un premier porte-outil (5) et un second porte-outil (6).	d-Et-un système de support réglable (10) qui supporte le siège (3) et qui est fixé à un point de fixation (9), la presse de frein (2) ayant un cadre (4) qui supporte un premier porte-outil (5) et un second porte-outil (6).
e-Le premier porte-outil (5) peut être déplacé par rapport au deuxième porte-outil (6) pour exécuter un mouvement de travail.	E-Le premier porte-outil (5) peut être déplacé par rapport au second support d'outil (6) pour effectuer un mouvement de travail.	e-Le premier porte-outil (5) peut être déplacé par rapport au deuxième porte-outil (6) pour exécuter un mouvement de travail.	E-Le premier porte-outil (5) peut être déplacé par rapport au second support d'outil (6) pour effectuer un mouvement de travail.
…	…	…	…

Table 2: structure of mixed translation texts.

Once created, we have started the series of assessments with our translator trainees as described in the following two sections.

2.2 First series of Assessment

As far as the first experiment is concerned, translator trainees were given three texts for each domain. The source document, patent and environmental discourse ones, its NTM version translation and its SMT version translation. The three texts were uploaded in Kantan, as shown in Figure 1 below, for the translator trainees to assess the adequacy and the fluency of each translation from a rank ranging from 1 to 5, with indeed a middle point

As already mentioned in section 2.1, no post-editing was possible in Kantan's comparison tool (called "A-B test"). Furthermore, translator trainees did know which translation was the result of the NMT version or of the SMT one.

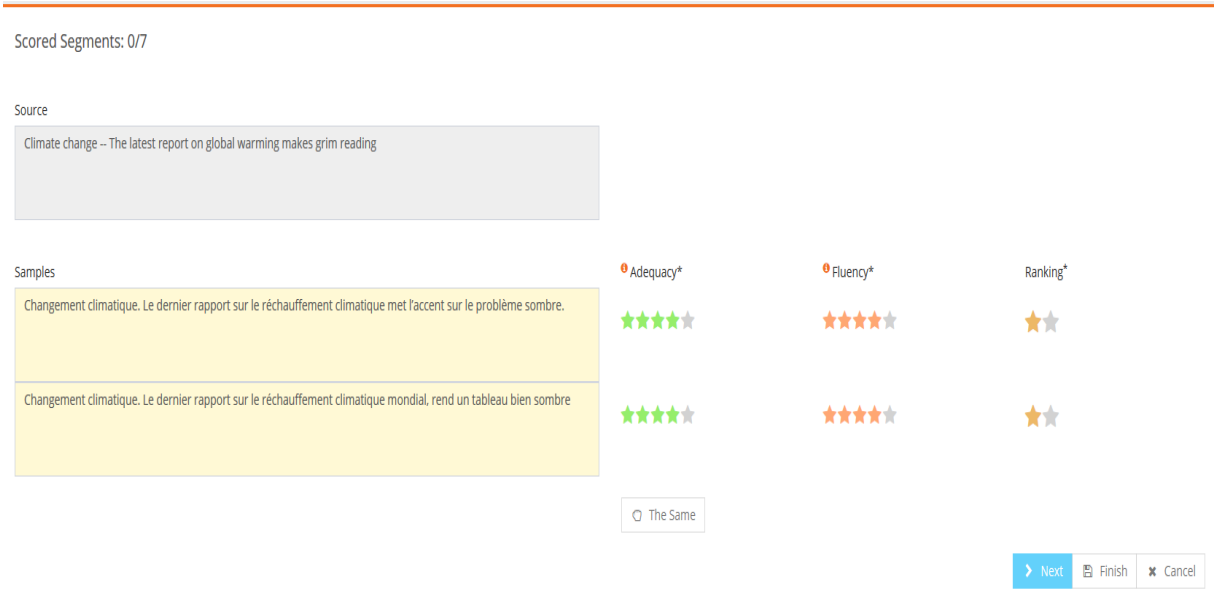

Figure 1. Kantan interface.

This first series of assessment were done to familiarise translator trainees to adequacy and fluency ranking. At the end of their course, trainees had to do the second series of assessment.

2.3 Second Series of Assessment

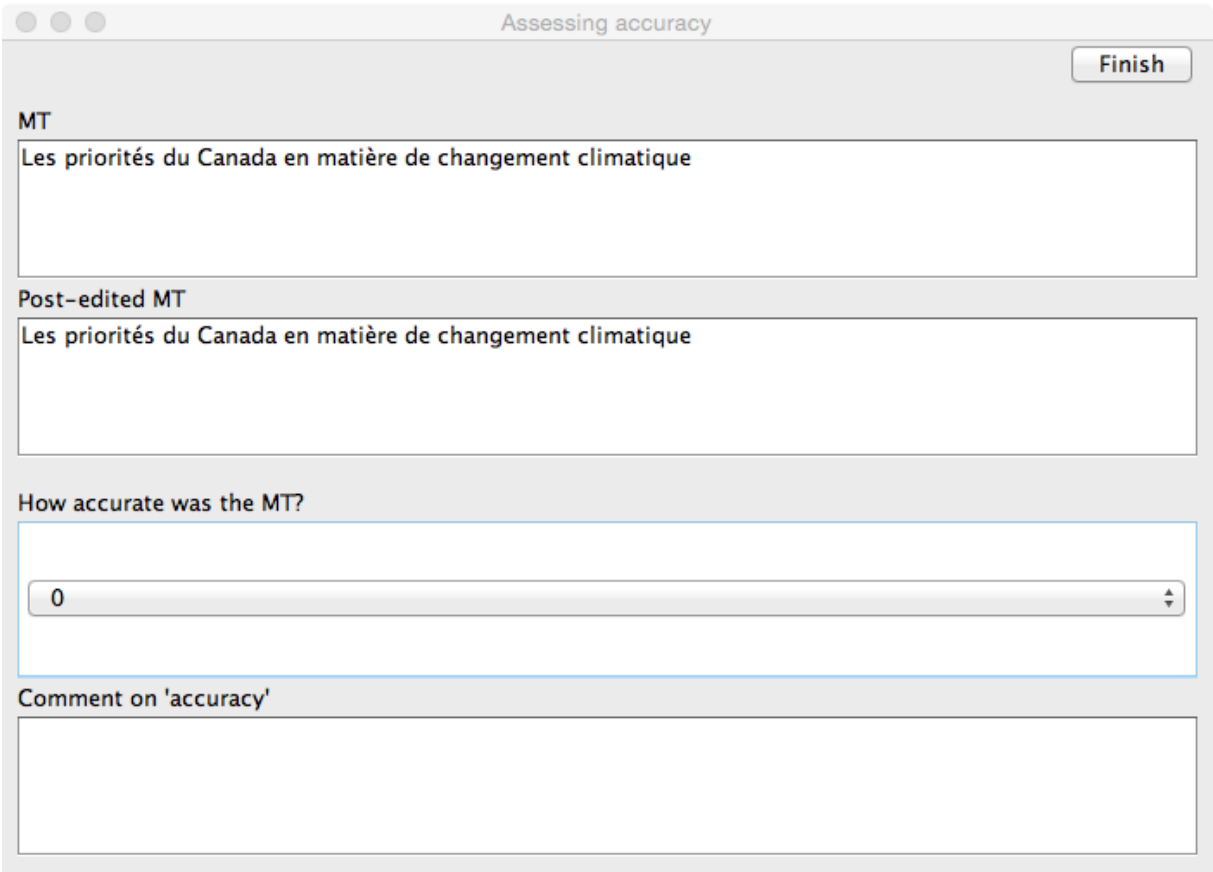

Figure 2. PET interface.

As said previously in section 2.1, for these series of assessment, the NMT and SMT versions were mixed into a new text, for the two domains, so that we would be sure the SMT and NMT versions would not be discovered by the translator trainees.

Four documents, one source document and one mixed-version translation document for each domain, were uploaded into PET. Before assessing the adequacy and fluency, the translator trainees had to post-edit the proposed translations.

After the post-edition, they were offered to score the accuracy and fluency with a rank ranging from 1 to 4, with currently no middle point, as can be seen in Figure 2.
In order to ensure the quality of the work, the translator trainees were presented this assessment as the final exam of their curriculum, thus accounting for their credits.

3 Results

Inter-rater reliability was assessed for each segment in each text using Fleiss' kappa for adequacy and fluency scores, and an intra class correlation coefficient (Vieira 2016: 52) for temporal measures. While the reliability of the measures collected without PE was low, the measures collected in PET were for the most part homogeneous.

3.1 Inter-rater reliability

In table 3, we compare overall agreement in percentage, together with Fleiss's Kappa. The latter does not appear to be fully relevant to our experimental setting. Indeed, we do not think that chance scoring is likely to happen at all in the context of a course on MT with MA students. The first piece of evidence for their involvement is the time spent assessing (see below table 4). Besides, the students had been trained to use adequacy and fluency scoring, and the two sessions considered for the present study were a training session for the final test, and the final test itself (in which students were given a mark for the course).

Overall, adequacy ratings in PET are the most reliable from the lowest 41.11% to the highest 60% opposed to Kantan adequacy rating ranging from 26.84% to 39.93%, which is much lower. Fluency ratings in PET are less reliable as they ranged from 30% to 50% (20 point raise), but still on average higher than Kantan results – the latter ranging from 27.29% to 39.65%. It appears that our students had more trouble assessing fluency with the patent extracts. Those results show that a post-editing task before assessing the adequacy of the translation provides a better agreement between raters.

According to domains, the post-edition of the environmental texts before assessing adequacy and fluency implied a better inter-rater agreement, respectively 26.84% vs. 41.11% (14 point raise) for SMT, 28.89% vs. 60% (31 point raise) for NMT and 32.65% vs. 50% (18 point raise) for SMT and 32.50% vs. 45.56% (13 point raise) for NMT. Nevertheless, it was less clear-cut for the post-edition of the patents before assessing adequacy, 39.93% vs. 46.67% (6.5 point raise) for SMT and 35.81% vs. 59.17% (23 point raise) for NMT, and fluency with 39.65% vs. 37.5% (2 point loss) for SMT and 27.29% vs. 30% (almost 3 point raise) for NMT. We could even notice with a loss of agreement of 2 points for the SMT version. This result can reinforce the fact that the translator trainees had more difficulties assessing fluency and Patents.

Concerning the different MT systems used, when comparing NMT vs. SMT, on the whole, we obtained a better agreement on adequacy for NMT than for SMT, 35.51% vs. 39.93% (4 point loss), 28.89% vs. 26.84% (2 point raise), 59.17% vs. 46.67% (12.5 point raise) and 60% vs. 41.11% (19 point raise). Again, the 4-point loss of agreement for NMT vs. SMT on Patents shows the difficulties of student to assess Patents. Moreover, we can see that the results are much better for the agreement on the adequacy for NMT when a post-edition had been performed prior to quality assessment. Contrastingly, the agreement on fluency, we obtained a better agreement for SMT than for NMT, 39.65% vs. 27.29% (12 point raise), 32.65% vs. 32.50% (equality), 37.6% vs. 30% (7 point raise), 50% vs 45.56% (4.5 point raise). We can see

that the difference among agreements as regards to fluency are less clear-cut, reinforcing the evidence that translator trainees had more difficulties to assess fluency. Once again, we can see that agreement is stronger when a post-edition had taken place before the quality assessment even if it is less obvious.

Data	**%Agreement (Fleiss's kappa) in adequacy rating of SMT**	**%Agreement (Fleiss's kappa) in adequacy rating of NMT**	**%Agreement (Fleiss's kappa) in fluency rating of SMT**	**%Agreement (Fleiss's kappa) in fluency rating of NMT**
Kantan Patents (WIPO Translate)	39.93% (0.05)	35.81% (0.08)	39.65% (0.12)	27.29% (0.06)
Kantan Climate (eTranslation)	26.84% (-0,01)	28.89% (0.07)	32.65% (0.06)	32.50% (0.11)
PET Patents (WIPO Translate)	46.67% (0.29)	59.17% (0.46)	37.50% (0.17)	30.00% (0.07)
PET Climate (eTranslation)	41.11% (0.05)	60.00% (0.14)	50.00% (0.33)	45.56% (0.27)

Table 3: Inter-rater Agreements

3.2 Intra class correlation (ICC) coefficients

Similarly, and even though there was more variation in temporal measures, homogeneity in ICC coefficients was stronger in PET data.

In Kantan, we could only get the overall time spent assessing both the SMT and NMT version translations for one source segment. We thus had to work on the duration means of assessing SMT and NMT in PET in order to be able to compare the results.

With Kantan, we can see in table 4 that there is remarkable variation from the Patent domain compared to the environmental domain. In contrast, PET has more homogeneous means, even though the standard deviation is very high and that the intra class correlation (ICC) also shows a lot of variation among raters. We can see that the post-editing time impacts the mean duration. We can also notice the the ICC seems to increase when less time is spent on assessing.

	Mean duration (ms) per sentence	**Standard Deviation**	**ICC**
Kantan Patents (WIPO Translate)	96.20558608	82.17102284	0.185130901
Kantan Climate (eTranslation)	249.3956044	271.804465	0.048161741
PET Patents (WIPO translate)	32193.90625	31266.08828	0.094569923
PET Climate (eTranslation	36006.725	45218.36249	0.05

Table 4: Intra class correlation coefficients

We finally sought to determine what went wrong by performing qualitative analyses of the problematic segments, as evidenced by both kappa and intra class correlation coefficients.

3.3 Qualitative Analyses

Qualitative analyses have been processed using the ACCOLÉ (Brunet-Manquat and Esperança-Rodier, 2018) annotation platform, as illustrated in Figure 3. Translation errors have been annotated according to DQF-MQM (Lommel and Melby, 2018) error typology and correlations between the different metrics have been calculated using Spearman's correlations.

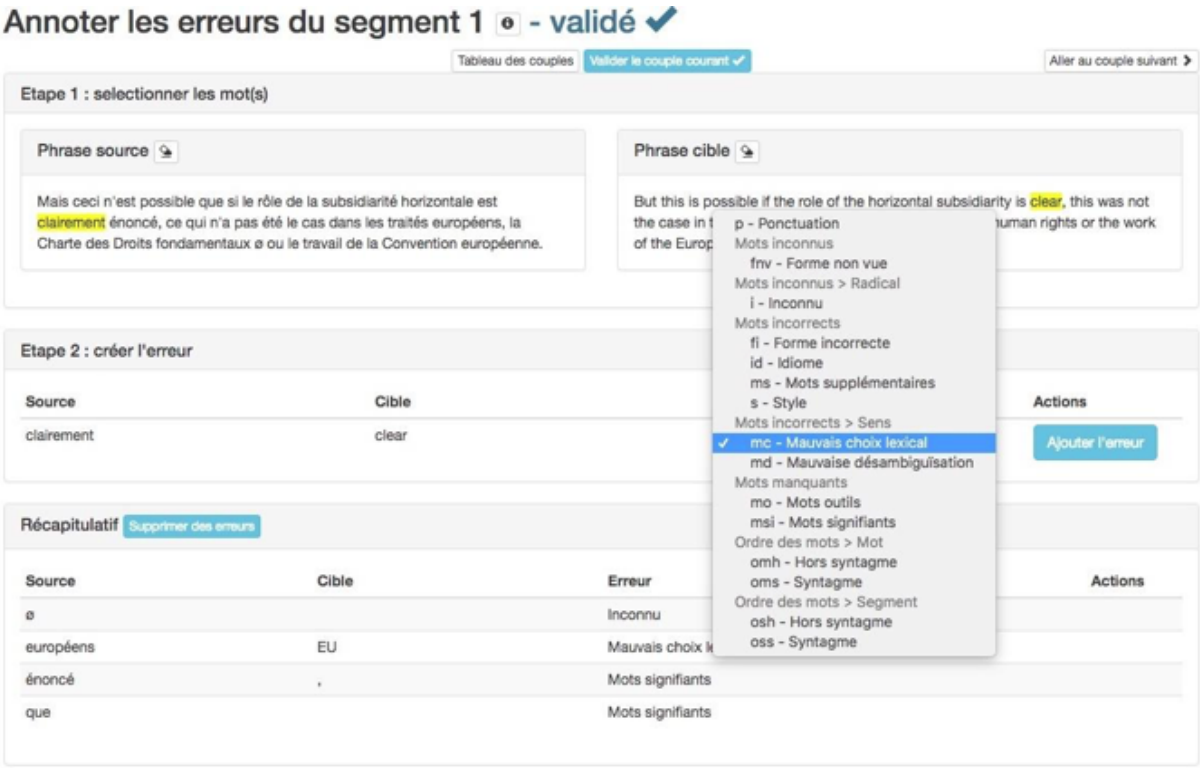

Figure 3. ACCOLÉ Interface.

We ended with the annotations of each sentence assessed by the trainee translators, according to the DQF-MQM typology.

		Time							
Time	r_s	1							
	p %	0	Hter						
hter	r_s	0.6103	1						
	p %	2.2525^{-7}	0	AssessingT					
Assessing T	r_s	0.4973	0.3885	1					
	p %	5.3038^{-5}	0.0021	0	Total errors				
Total errors	r_s	0.3599	0.2104	0.056	1				
	p %	0.0047	0.1066	0.6707	0	Accuracy E			
Accuracy E	r_s	0.3991	0.1971	0.1501	0.9852	1			
	p %	0.0016	0.1311	0.2523	4.0900^{-46}	0	Fluency E		
Fluency E	r_s	0.2634	0.0915	0.8123	0.8575	0.7833	1		
	p %	0.0420	0.4867	3.3627^{-15}	2.2014^{-18}	1.3894^{-13}	0	Adequacy	
Adequacy	r_s	-0.481	-0.5693	-0.3785	-0.2904	-0.2879	-0.2091	1	
	p %	0.0001	2.0636^{-6}	0.0029	0.0244	0.0257	0.1088	0	Fluency
Fluency	r_s	-0.5208	-0.6071	-0.4775	0.0237	0.0021	0.1659	0.4167	1
	p %	1.9916^{-5}	2.7045^{-7}	0.0001	0.8574	0.9873	0.2052	0.0009	0

Table 5: Spearman's correlation - NMT translation - environmental domain.

For easier understanding, we have counted for each sentence, on the one hand, the number of annotated errors corresponding to the accuracy type and on the other hand, the number of annotated errors corresponding to the fluency type.

Table 5 shows the Spearman's correlation calculation done for each measure obtained for the NMT translation version we had for the environmental domain.

Time is strongly correlated to HTER (r_s=0.6103 with p-value=2.2525^{-7}) as well as the assessing time (r_s=0.4973 with p-value=5.3038^{-5}) which leads us to say that time spent is dedicated to the post-edition. This is confirmed by the fact that HTER is correlated to the Assessing time (r_s=0.3885 with p-value=0.0021).

Furthermore, the Assessing time is highly correlated to the presence of Fluency errors (r_s=0.8123 with p-value=3.3627^{-15}). Consequently, we can say that when there are Fluency errors the assessing time is longer.

As regards the correlation between Adequacy and Fluency (r_s=0.4167 with p-value=0.0009), we can see that the scores given to the Adequacy of a translation have an influence on the score given to Fluency. In the same way, looking at the correlation between Accuracy errors and Fluency errors (r_s=0.7830 with p-value=1.3894^{-13}), we can see that there is also a link leading us to say that the scores given to the Accuracy errors of a translation have an influence on the score given to the Fluency errors.

Logically enough, the correlation between the Total errors and the Accuracy errors and the Fluency Errors is really high.

It is also interesting to focus on the strong negative correlations from Adequacy and Fluency as regards Time, Hter and Assessing Time. Those negative correlations mean that, when there is a lot of time spent, and that there are a lot of changes between the translation and the reference, then the Adequacy and Fluency scores are much lower, and conversely.

		Time							
Time	r_s	1							
	p %	0	Hter						
hter	r_s	0.288	1						
	p %	0.0256	0	AssessingT					
Assessing T	r_s	0.2568	0.3423	1					
	p %	0.0479	0.0074	0	Total errors				
Total errors	r_s	0.2715	0.0841	0.0761	1				
	p %	0.0358	0.5228	0.5634	0	Accuracy E			
Accuracy E	r_s	0.3698	0.0474	0.186	0.687	1			
	p %	0.0036	0.7193	0.1547	1.3528^{-9}	0	Fluency E		
Fluency E	r_s	0.0965	-0.0009	-0.09	0.7656	0.0955	1		
	p %	0.4635	0.9946	0.49442	1.0415^{-12}	0.4681	0	Adequacy	
Adequacy	r_s	-0.2082	-0.4452	-0.4565	-0.2578	-0.1758	-0.097	1	
	p %	0.1103	0.0003	0.0002	0.0467	0.1791	0.4608	0	Fluency
Fluency	r_s	-0.5046	-0.4432	-0.3071	-0.1798	-0.1025	-0.1553	0.4428	1
	p %	3.9331	0.0004	0.0170	0.1691	0.4359	0.2360	0.0004	0

Table 6: Spearman's correlation - SMT translation - environmental domain.

Table 6 shows Spearman's correlation for measures taken for the SMT translation on the same domain as the one for NMT.

We almost find the same correlations, nevertheless, the correlations between Time and Hter (r_s=0.3423 with p-value=0.0074) is weaker for the SMT translation opposed to NMT translations. The lower value suggests that there is more variation and that there might be more effort, as captured by variations in Time measurements, while Hter scores remain relatively similar. It is also the case for the correlations as regards Hter and Adequacy as well as Fluency, which are still strong negative correlations. As was the case for NMT, but in a smaller

proportion, when the Hter score is high, the scores for Adequacy and Fluency are low. The lower proportion suggests that students might be less good at using Adequacy and Fluency measures for assessing effort with SMT.

Again, and as expected, the correlations between Total errors and Accuracy errors and Fluency errors are really strong.

Finally, for the SMT translation, the Assessing time is strongly correlated to Adequacy scores while for the NMT translation it was correlated to Fluency Scores. This statement matches previous NMT vs. SMT quality assessments showing that NMT translations are more fluent than SMT ouputs, while SMT translations are more adequate (see e.g. Koehn & Knowles 2017).

4 Discussion

Results showed strong correlations, whether positive or negative, between time spent post-editing and all the other metrics for NMT but weakest correlations as far as SMT was concerned. Our results thus point to much more homogeneity in post-editing NMT outputs, with more variation in the treatment of SMT errors. The consequences of these differences for professional post-editors include both lighter cognitive effort and improved cognitive ergonomics when dealing with NMT. It is also worth noting that estimated effort (as expressed by Fluency and Adequacy ratings) was on the whole more realistic with NMT outputs. However, it remains to be seen how post-editors will address the higher risks induced by more homogeneous and fluent NMT outputs, notably that of meaning errors going unnoticed (Forcada, 2017: 303). Experimental designs including hidden errors and allowing for a measure of cognitive effort would help in determining whether the necessary attention to details with NMT outputs ends up being more demanding than the very regular, and somewhat teadious, post-editing tasks required with SMT.

Acknowledgements

The work reported above has been granted by NeuroCoG/Pôle Grenoble Cognition funding as well as LIG/Emergence funding.

References

Aziz, Wilker, Sheila Castillo Maria de Sousa, and Lucia Specia. 2012. PET: a Tool for Post-editing and Assessing Machine Translation. In *Proceedings of the 16th Annual Conference of the European Association for Machine Translation*, pages 3982-3987.

Brunet-Manquat, Francis and Emmanuelle, Esperança-Rodier. 2018. ACCOLÉ : Annotation Collaborative d'erreurs de traduction pour COrpus aLignÉs . DEMONSTRATION, in *Proceedings of the 25e conférence sur le Traitement Automatique des Langues Naturelles (TALN),* Rennes.

Castilho, Sheila, Stephen Doherty, Federuci Gaspari and Joss Moorkens. 2018. Approaches to human and machine translation quality assessment. In *Translation Quality Assessment*. Pages 9-38. Springer, Cham.

Forcada, Michael L. 2017. Making sense of neural machine translation. In *Translation spaces*, 6(2), pages 291-309.

Koehn, Philipp and Rebecca Knowles. 2017. Six challenges for neural machine translation. arXiv preprint arXiv:1706.03872.

Lommel, Arle and Alan K. Melby. 2018. Tutorial: MQM-DQF: A Good Marriage (Translation Quality for the 21st Century). *In proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 2: User Papers)*. Vol. 2.

Vieira, Lucas N. 2016. How do measures of cognitive effort relate to each other? A multivariate analysis of post-editing process data. *Machine Translation*, 30(1-2), pages 41-62.

Way, Andy. 2018a. Machine translation: where are we at today? In *Angelone E, Massey G, Ehrensberger-Dow M (eds) The Bloomsbury companion to language industry studies*. Bloomsbury, London.

Way, Andy. 2018b. Quality expectations of machine translation. In *Translation Quality Assessment.* pages 159-178. Springer, Cham.