Post-editese: an Exacerbated Translationese

Antonio Toral
Center for Language and Cognition
University of Groningen
The Netherlands
a.toral.ruiz@rug.nl

Abstract

Post-editing (PE) machine translation (MT) is widely used for dissemination because it leads to higher productivity than human translation from scratch (HT). In addition, PE translations are found to be of equal or better quality than HTs. However, most such studies measure quality solely as the number of errors. We conduct a set of computational analyses in which we compare PE against HT on three different datasets that cover five translation directions with measures that address different translation universals and laws of translation: simplification, normalisation and interference. We find out that PEs are simpler and more normalised and have a higher degree of interference from the source language than HTs.

1 Introduction

Machine translation (MT) is nowadays widely used in industry for dissemination purposes by means of post-editing (PE, also referred to as PEMT in the literature), a machine-assisted approach to translation that results in notable increases in translation productivity compared to unaided human translation (HT), as shown in numerous research studies, e.g. Plitt and Masselot (2010).

In theory, one would claim that HTs[1] and PE translations are clearly different, since, in the translation workflow of the latter, the translator is primed by the output of an MT system (Green et al., 2013), resulting in a translation that should then contain, to some extent, the footprint of that MT system. Because of this, one would conclude that HT should be preferred over PE, as the former should be more natural and adhere more closely to the norms of the target language. However, many research studies have shown that the quality of PE is comparable to that of HT or even better, e.g. Koponen (2016), and, according to one study (Bowker and Buitrago Ciro, 2015), native speakers do not have a clear preference for HT over PE.

In this paper we conduct a set of computational analyses on several datasets that contain HTs and PEs, involving different language directions and domains as well as PE performed according to different guidelines (e.g. full versus light). Our aim is to find out whether HT and PE differ significantly in terms of different phenomena. Since previous research has proven the existence of translationese, i.e. the fact that HT and original text exhibit different characteristics, our current research can be framed as a quest to find out whether there is evidence of *post-editese*, i.e. the fact that HT (or translationese) and PE would be different.

The characteristics of translationese can be grouped along the so-called universal features of translation or translation universals (Baker, 1993), namely simplification, normalisation (also referred to as homogeneisation) and explicitation. In addition to these three, interference is recognised as a fundamental law of translation (Toury, 2012): "phenomena pertaining to the make-up of the source text tend to be transferred to the target text". In a nutshell, compared to original texts, translations tend to be simpler, more standardised, and

[1]By HT we refer to translations produced by a human from scratch, i.e. without the assistance of an MT system or any other computer-assisted technology, e.g. translation memories.

more explicit and they retain some characteristics that pertain to the source language.

In this study then we study the existence of post-editese by conducting a set of computational analyses that fall into three[2] out of these four categories. With these analyses we aim to answer a number of research questions:

- **RQ1**. Does post-editese exist? I.e. is there evidence that PE exhibits different characteristics than HT?
- **RQ2**. If the answer to RQ1 is yes, then which are the main characteristics of PE? I.e. how does it differ from HT?
- **RQ3**. If the answer to RQ1 is yes, then, are there different post-editeses? I.e. are there any characteristics that distinguish the post-editese produced by MT systems that follow different paradigms (rule-based, statistical phrase-based and neural)?

The rest of this paper is organised as follows. Section 2 provides an overview of the related work. Section 3 covers the experimental setup and the experiments conducted. Finally, Section 4 presents our conclusions and suggests lines of future work.

2 Related Work

Many research studies carried out during the last decade have compared the quality of HT and PE. These have shown that the quality of PE is comparable to that of HT, e.g. García (2010), or even better, e.g. Guerberof (2009) and Plitt and Masselot (2010). In these studies quality is typically measured in terms of the number of mistakes in each translation condition. However, they do not take into account other relevant aspects that may flag important differences between HTs and PEs, such as the perspective of the end-user or the presence of different phenomena in both types of translations. A recent strand of work targets precisely these, (i) by collecting the preference of end users between HT and PE and (ii) by analysing the characteristics of both types of translations. In the following subsections we report on recent work conducted in both of these two research lines.

2.1 Preference between HT and PE

Fiederer and O'Brien (2009) compared HT and PE for the English-to-German translation of text from a software user manual. Participants ranked both conditions equally for clarity, PE higher than HT for accuracy and HT higher than PE for style. When asked to choose their favourite translation, HT obtained a higher percentage of preferences than PE: 63% to 37%.

Bowker (2009) presented French- and English-speaking minorities in Canada four translations (HT, full PE, light PE and raw MT) of three short governmental texts (approximately 325 words) written in a relatively clear, neutral style with reasonably short sentences. Preference took into account not only the quality of the translations but also the time and cost required to produce them. The results were rather different for each group. In the French-speaking minority group 71% preferred HT versus 29% PE (21% full and 8% light). However, half of the participants were language professionals, which skewed the results. In fact, when they were removed the results changed drastically: 56% preferred HT versus 44% PE (29% full, 15% light). As for the English minority, 8% preferred HT versus 87% PE (38% full, 49% light) and 4% raw MT.

Bowker and Buitrago Ciro (2015) presented Spanish-speaking immigrants in Canada with four translations (HT, full PE, light PE and raw MT) of three short texts (301 to 380 words) containing library-related information and asked them which they prefered. PE and HT attained a similar number of preferences; 49% of respondents preferred PE (24% full and 25% light, respectively) compared to 42% who preferred HT. Raw MT lagged considerably behind with the remaining 9% of the preferences.

Green et al. (2013) assessed the quality of HT versus PE for Wikipedia articles translated from English into Arabic, French and German. Quality was measured by means of preference, done by ranking on isolated sentences via crowdsourcing. PE was found significantly better for all translation directions.

2.2 Characteristics of HT and PE

Daems et al (2017) used HT and PE news translated from English into Dutch. They presented them to translation students and colleagues, whose task was to identify which translations were PE. They also tried an automated approach, for which they built a classifier with 55 features using surface forms and linguistic information at lexical, syntac-

[2] Explicitation is not addressed in our experiments.

Dataset	Direction	PE type	MT systems	# Sent. pairs	Domain
Taraxü	en→de de→en es→de[4]	Light[3]	2 SMT, 2 RBMT	272 240 101	News
IWSLT	en→de en→fr	Light	4 NMT, 4 SMT 2 NMT, 3 SMT	600	Subtitles
MS	zh→en	Full	1 NMT[5]	1,000[6]	News

Table 1: Information about the datasets used in the experiments

tic and semantic levels. No proof of the existence of post-editese was found, either perceived (students) or measurable (classifier).

Čulo and Nitzke (2016) compared MT, PE and HT in terms of terminology and found that the way terminology is used in PE is closer to MT than to HT and has less variation than in HT.

Farrell (2018) identified MT markers (i.e. "translation solutions which occurred with a statistically significantly higher frequency in PEMT than in HT") in short texts from Wikipedia translated from English into Italian and found that MT tends to choose a subset of all the possible translation solutions (the most frequent ones) and that this is the case also, to some extent, in PEs. HTs and PEs were also compared in terms of number of errors, which were found to be comparable, corroborating the findings of the literature covered at the beginning of this section.

Our contribution falls into this research line, to which we contribute a computational study whose analyses are chosen to align to translation universals and laws of translation and that covers multiple languages and domains.

3 Experiments

In this section we first describe the datasets used (Section 3.1), and then report on each of the experiments that we carried out in the subsequent subsections: lexical variety (Section 3.2), lexical density (Section 3.3), length ratio (Section 3.4) and part-of-speech sequences (Section 3.5).

3.1 Datasets

We make use of three datasets in all our experiments: Taraxü (Avramidis et al., 2014), IWSLT (Cettolo et al., 2015; Mauro et al., 2016) and Microsoft "Human Parity" (Hassan et al., 2018), henceforth referred to as MS. These datasets cover five different translation directions that involve five languages:[7] English↔German, English→French, Spanish→German and Chinese→English. In addition, this choice of datasets allows us to include a longitudinal aspect into the analyses since there are state-of-the-art MT systems from almost one decade ago (in Taraxü), from three and four years ago (IWSLT) and from just one year ago (MS). Table 1 shows detailed information about each dataset, namely its translation direction(s), type of PE done, paradigm of the MT system(s) used, number of sentence pairs and domain of its text.

We note the following two limitations in some of the datasets:

- Mismatch of translator competence. Both PE and HT are carried out by professional translators in two of the datasets (Taraxü and MS). However, in the remaining one, IWSLT, professional translators do PE, while the translators doing HT are not necessarily professionals[8]. Thus, if we find differences between PEs and HTs, for this dataset this may not be (entirely) due to the two translations procedures leading to different translations but (also) to the different translations being produced by translators with different levels of proficiency.

[3] "Translators were asked to perform only the minimal post-editing necessary to achieve an acceptable translation quality." (Avramidis et al., 2014)

[4] This dataset contains an additional translation direction (de→es) which is not used here due to its small size; 40 sentence pairs.

[5] The MT systems used in the MT and in the PE condition are not the same. The one in the MT condition is the best system in Hassan et al. (2018) while the one in the PE condition is Google Translate, again as provided in Hassan et al. (2018).

[6] The original dataset contains 2,001 sentences. We only use the subset in which the source text is original instead of translationese (Toral et al., 2018).

[7] In the tables and experiments we will refer to languages with their ISO-2 codes.

[8] "We accept all fluently bilingual volunteers as translators", `https://translations.ted.com/TED_Translator_Resources:_Main_guide#Translation`

Translation type	Dataset and translation direction					
	Taraxü			IWSLT		MS
	de→en	en→de	es→de	en→de	en→fr	zh→en
HT	0.26	0.27	0.31	0.20	0.16	0.14
PE	**-2.05%**	**-1.81%**	†**-1.27%**	**-3.86%**	**-1.17%**	**-4.76%**
MT	-2.94%	-3.62%	-5.91%	-10.93%	-6.04%	-6.96%
PE-NMT				-4.21%	-1.88%	-4.76%
PE-SMT	**-1.59%**	**-1.31%**	†**-1.03%**	**-3.50%**	**-0.70%**	
PE-RBMT	-2.79%	-2.04%	-3.05%			
NMT				-12.22%	-8.18%	-7.33%
SMT	**-2.36%**	**-2.36%**	**-6.42%**	**-9.63%**	**-4.61%**	
RBMT	-3.08%	-4.26%	-7.78%			

Table 2: TTR scores for HT and relative differences for PE and MT. For directions with more than one MT system, the result shown in rows PE and MT uses the average score of all the PEs or MT outputs, respectively. The best result (highest TTR) in each group of rows is shown in bold. If a † is not shown then the TTR for HT is significantly higher than the TTRs for all the translations in that cell (the 95% confidence interval of the TTR of HT, obtained with bootstrap resampling, is higher and there is no overlap).

- Source language being translationese. In two of the datasets (MS and IWSLT), the source language and the language in which those texts were originally written is the same. This is not the case however for Taraxü, for which the original language of the source texts is Czech. We can still compare MT to PE although we need to take into account that these texts are easier for MT than original texts (Toral et al., 2018). However the comparison between PE (or MT) and HT is problematic since the HT was not translated from the source language but from another language (Czech).

3.2 Lexical Variety

We assess the lexical variety of a translation (MT, PE or HT) by calculating its type-token ratio (TTR), as shown in equation 1.

$$TTR = \frac{number\ of\ types}{number\ of\ tokens} \quad (1)$$

Farrell (2018) observed that MT tends to produce a subset of all the possible translations in the target language (the ones used most frequently in the training data). Therefore, we hypothesise the TTR of MT, and by extension that of PE too, to be lower than that of HT. If this is the case, then PE would be, in terms of lexical variety, simpler than HT.

Table 2 shows the results for each dataset and language direction. In all cases the lexical variety in PE is lower than in HT, and again in all cases, that of MT is lower than that of PE. This could be interpreted as follows: (i) lexical variety is low in MT because these systems prefer the translation solutions most frequently used in the training data; (ii) a post-editor will add lexical variety to some degree, but because MT primes him/her, the resulting translation will not achieve the level of lexical diversity that is attained in HT.

We now look at the results of PE and MT for different MT paradigms. In the Taraxü dataset we can compare rule-based and statistical MT systems. Rule-based MT has a lower TTR in all three translation directions of this dataset and this is then reflected in a lower TTR again when a system of this paradigm is used for post-editing. In the IWSLT dataset we can confront statistical and neural MT systems. In all cases the lexical variety of neural MT is lower than that of statistical MT. Again, the same trend shows when we look at their PEs. This is perhaps a surprising result, since NMT systems outperformed SMT in the IWSLT dataset, in terms of HTER (Cettolo et al., 2015).

3.3 Lexical Density

Lexical density measures the amount of information present in a text by calculating the ratio between the number of content words (adverbs, adjectives, nouns and verbs) and its total number of words, as shown in equation 2.

$$lex\_density = \frac{number\ of\ content\ words}{number\ of\ total\ words} \quad (2)$$

Translationese has been found to have a lower percentage of content words than original texts,

Translation Type	Dataset and translation direction					
	Taraxü			IWSLT		MS
	de→en	en→de	es→de	en→de	en→fr	zh→en
HT	0.55	0.53	0.53	0.48	0.46	0.59
PE	-1.00%	-2.48%	**-4.31%**	**-3.46%**	-1.24%	**-0.46%**
MT	**-0.81%**	**-0.69%**	-4.53%	-5.14%	**-0.94%**	-2.37%
PE-NMT				-3.88%	-1.47%	-0.46%
PE-SMT	**-0.54%**	-2.87%	-4.78%	**-3.04%**	**-1.09%**	
PE-RBMT	-1.46%	**-2.09%**	**-3.84%**			
NMT				-6.31%	-3.14%	-2.37%
SMT	**-0.80%**	**0.14%**	**-3.45%**	**-3.98%**	**0.53%**	
RBMT	-0.83%	-1.51%	-5.61%			

Table 3: Lexical density scores for HT and relative differences for PE and MT. For directions with more than one MT system, the result shown in rows PE and MT uses the average score of all the PEs or MT outputs, respectively. The best result (highest density) in each group of rows is shown in bold.

thus being, from this point of view, lexically simpler (Scarpa, 2006). To identify and count content words we tag the target sides of the datasets with their parts-of-speech (PoS) using UDPipe (Straka et al., 2016), a PoS tagger that uses the Universal PoS tagset.[9] Then we assess the lexical density of each translation (HT, PE and MT) using this PoS-tagged version[10] of the datasets.

Table 3 shows the results. In both PE and MT the lexical density is lower than in HT. However between PE and MT, there is no systematic distinction. When inspecting PEs using different MT paradigms, we do not find any clear trend between SMT and RBMT, but one such trend shows up when we inspect SMT and NMT: in the two comparisons we can establish in our dataset (the two translation directions in the IWSLT dataset), PE-NMT leads to a lower lexical density than PE-SMT. Finally, looking at MT outputs produced by different types of MT systems, we observe that both RBMT and NMT lead to lower lexical densities than SMT.

3.4 Length Ratio

Given a source text ST and a target text TT, i.e. TT is a translation of the ST (HT, PE or MT), we compute the absolute difference in length (measured in characters) between the two, normalised by the length of the ST, as shown in equation 3.

$$length\ ratio = \frac{|length_{ST} - length_{TT}|}{length_{ST}} \quad (3)$$

Because (i) MT results in a translation of similar length to that of the ST,[11] and PE is primed by the MT output while (ii) a translator working from scratch (HT) may translate more freely in terms of length, we hypothesise that the difference in absolute length is smaller for MT and PE than it is for HT. If this is true, it would be a case of interference in PE, as the typical length of sentences translated with this method would be similar to the length used in the source text.

We compute this ratio at sentence level and average over all the sentences of the dataset. Table 4 shows the results for each dataset and language direction. The results in datasets Taraxü and MS match our hypothesis; in both datasets the length ratio is lower for PE and MT than it is for HT. This is also the case for MT in dataset IWSLT. However, in the results for PE in dataset IWSLT, the ratio of PE is actually higher than that of HT for en→fr, which seems to contradict our hypothesis. This may be attributed to the difference in translation proficiency between the translators that did HT and those that did PE that we commented upon in Section 3.1. The latter are professional, while the first could be non-professional. It is known

[9] `https://universaldependencies.org/u/pos/`
[10] UDPipe's PoS tagging F1-score is over 90% for all the three target languages considered: de, en and fr (Straka and Straková, 2017, Table 2)

[11] This is necessary the case for RBMT and SMT as the number of TL tokens they can produce per each SL token is limited; e.g. the longest a translation with SMT can be is the number of tokens in the ST multiplied by the longest phrase in the phrase table, which is typically 7. NMT does not have this limitation, so we do not argue in this direction for that MT paradigm.

Dataset	Direct.	Length ratio		
		HT	PE	MT
Taraxü	de→en	0.16	‡-38.5%	†-36.9%
	en→de	0.22	‡-33.4%	‡-38.5%
	es→de	0.17	*-25.2%	-21.0%
IWSLT	en→de	0.17	-3.4%	†-18.8%
	en→fr	0.18	6.7%	-10.9%
MS	zh→en	1.41	‡-9.9%	‡-9.1%

Table 4: Length ratio scores for HT and relative differences for PE and MT. For directions with more than one MT system, the result shown in columns PE and MT uses the average length ratio of all the PEs or MT outputs, respectively. * indicates that the score for HT is significantly higher with $\alpha = 0.05$ († with $\alpha = 0.01$ and ‡ with $\alpha = 0.001$) than the scores of all the PEs/MTs represented in the cell, based on one-tailed paired t-tests.

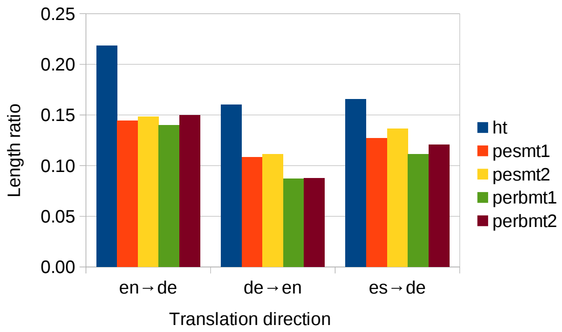

Figure 1: Length ratio for HT and PEs in the Taraxü dataset.

that non-professional translators tend to produce more literal translations, whose length should then be similar to that of the source text.

We now look at the length ratio of PEs that use different MT systems. Figure 1 shows the length ratios of HTs and PEs that use different MT paradigms (SMT and RBMT) in the Taraxü dataset. While for one of the translation direction (en→de) the ratio of PE-SMT and PE-RBMT are similar, the two PE-RBMT systems have lower length ratios than the two PE-SMT systems in the other two translation directions.

3.5 Part-of-Speech Sequences

We assess the interference of the source language on a translation (HT, PE and MT) by measuring how close the sequence of PoS tags in the translation is to the typical PoS sequences of the source language and to the typical PoS sequences of the target language. If the sequences of PoS tags used in a translation A are more similar to the typical sequences of the source language than the sequences of another translation B, that could be an indication that A has more interference from the source language than B.

Given a PoS-tagged translation T, a language model of a PoS-tagged corpus in the source language LM_{SL} and a language model of a PoS-tagged corpus in the target language LM_{TL}, we calculate the difference of the perplexities of T with respect to both language models, as shown in equation 4.

$$PP\_diff = PP(T, LM_{SL}) - PP(T, LM_{TL}) \tag{4}$$

A high result for a translation T would indicate that T is dissimilar to the source language (high perplexity with respect to the source language) and similar to the target language (low perplexity with respect to the target language). Conversely, a low result would indicate that T is similar to the source language (low perplexity with respect to the source language) and dissimilar to the target language (high perplexity with respect to the target language).

Because MT systems are known to perform less reordering than human translators (Toral and Sánchez-Cartagena, 2017), our hypothesis is that MT outputs, and by extension PEs, are more similar in terms of PoS sequences to the source language than HTs are. This would mean that PEs and MT outputs have more interference in terms of PoS sequences than HTs.

For each language in our datasets (de, en, es, fr and zh), we PoS tag a monolingual corpus[12] with UDPipe, the PoS tagger already introduced in Section 3.3.[13]

We then build language models on these PoS tagged data with SRILM (Stolcke, 2002), considering n-grams up to $n = 6$, using interpolation and Witten-Bell smoothing.[14] Because we use the Uni-

[12]The corpora used for the different languages belong to the same domain, news. We use corpora of roughly the same size (around 100MB of text). This leads to corpora of between 1,617,527 sentences (es) and 2,187,421 (de).

[13]While in Section 3.3 we PoS-tagged the target side of the datasets, in this experiment we PoS-tag both the source and target sides. UDPipe's PoS tagging F1-score is over 90% for four of the five languages involved (de, en, es and fr) and 83% for the remaining one: zh (Straka and Straková, 2017, Table 2). Given the lower performance of PoS tagging for zh, the results involving this language should be taken with caution.

[14]The more advanced smoothing method Kneser-Ney did not work because the count-of-count statistics in our datasets are not suitable for this smoothing method, which may be due to the very small size of our vocabulary: the Universal Dependencies PoS tagset.

Translation Type	Dataset and translation direction					
	Taraxü			IWSLT		MS
	de→en	en→de	es→de	en→de	en→fr	zh→en
HT	5.12	5.09	9.41	5.01	2.47	17.23
PE	**-13.84%**	**-11.29%**	**-8.58%**	**-6.26%**	**-2.03%**	-3.26%
MT	-33.65%	-32.25%	-20.71%	-18.66%	-11.07%	**-3.1%**
PE-NMT				**-3.41%**	**-1.40%**	-3.26%
PE-SMT	**-11.72%**	-13.37%	-10.48%	-9.10%	-2.46%	
PE-RBMT	-15.95%	**-9.20%**	**-6.68%**			
NMT				**-5.89%**	**-2.58%**	-3.10%
SMT	**-30.07%**	-41.71%	-26.30%	-31.43%	-7.95%	
RBMT	-37.24%	**-22.80%**	**-15.13%**			

Table 5: Perplexity difference scores for HT and relative differences for PE and MT. For directions with more than one MT system, the result shown in rows PE and MT uses the average score of all the PEs or MT outputs, respectively. The best result (highest perplexity) in each group of rows is shown in bold.

versal PoS tagset, the set of PoS tags is the same for all the languages, which means that all our language models share the same vocabulary.

Table 5 shows the results. In terms of HT versus PE and MT, we observe similar trends to those observed earlier for lexical variety (Table 2), namely MT obtains the lowest perplexity difference score and HT the highest, with PE lying somewhere between the two. The only exception to this is seen in the MS dataset, where the value for MT is slightly higher than that for PE (-3.1% versus -3.26%). It should be taken into account, as already explained in Section 3.1, that the MT systems in the MT and PE conditions are different in this dataset, with the one in the MT condition being substantially better (Hassan et al., 2018). Overall, we interpret these results as MT being the translation type that contains the most interference in terms of PoS sequences, followed by PE.

We now look at the PE and MT results under different MT paradigms. Comparing SMT and NMT, the results indicate that the latter has less interference, both in PE and MT conditions. This corroborates earlier research that compared SMT and NMT in terms of reordering (Bentivogli et al., 2016). We do not find clear trends when comparing SMT and RBMT though.

4 Conclusions and Future Work

We have carried out a set of computational analyses on three datasets that contain five translation directions with the aim of finding out whether post-edited translations (PEs) exhibit different phenomena than human translations from scratch (HTs). In other words, whether there is evidence of the existence of *post-editese*. The analyses conducted measure aspects related to translation universals and laws of translation, namely simplification, normalisation and interference. With these analysis, we find evidence of post-editese (RQ1), whose main characteristics (RQ2) we summarise as follows:

- PEs have lower lexical variety and lower lexical density than HTs. We link these to the simplification principle of translationese. Thus, these results indicate that *post-editese* is lexically simpler than translationese.

- Sentence length in PEs is more similar to the sentence length of the source texts, than sentence length in HTs. We link this finding to interference and normalisation: (i) PEs have interference from the source text in terms of length, which leads to translations that follow the typical sentence length of the source language; (ii) this results in a target text whose length tends to become normalised.

- Part-of-speech (PoS) sequences in PEs are more similar to the typical PoS sequences of the source language, than PoS sequences in HTs. We link this to the interference principle: the sequences of grammatical units in PEs preserve to some extent the sequences that are typical of the source language.

In the paper we have not considered only HTs and PEs but also MT outputs, from the MT systems that were the starting point to produce the PEs. This to corroborate a claim in the literature (Green et al., 2013), namely that in PE the translator is primed by the MT output. We expected then to find similar trends to those found in PEs also in MT outputs and this was indeed the case in all four experiments. In two of the experiments, lexical variety and PoS sequences, the results of PE were somewhere in the middle between those of HT and MT. Our interpretation is that a post-editor improves the initial MT output in terms of variety and PoS sequences, but due to being primed by the MT output, the result cannot attain the level of HT, and the footprint of the MT system remains in the resulting PE.

We have also looked at different MT paradigms (rule-based, statistical and neural), to find out whether these lead to different characteristics in the resulting PEs (RQ3). Neural MT diminishes to some extent the interference in terms of PoS sequences, probably because it is better at reordering (Bentivogli et al., 2016). Statistical MT has obtained better results than the other two MT paradigms in terms of lexical variety, and also better than neural MT, but on par with rule-based MT, in lexical density.

In a nutshell, we have found that PEs tend to be simpler and more normalised and to have a higher degree of interference from the source text than HTs. This seems to be caused because these characteristics are already present in the MT outputs that are the starting point of the PEs. We find thus evidence of *post-editese*, which can be thought of as an exacerbated translationese.

While PE is very useful in terms of productivity, which arguably is the main reason behind its wide adoption in industry, the findings of this paper flag a potential issue. Because PEs are simpler and have a higher interference from the source language than HTs, the extensive use of PE rather than HT may have serious implications for the target language in the long term, for example that it becomes impoverished (simplification) and overly influenced by the source language (interference). At the same time, we have shown that these issues cannot be attributed to PE *per se* but that they originate in the MT systems used as the starting point for PE. Identifying these issues might be then the first step for further research on addressing these problems in current state-of-the-art MT systems.

Throughout the paper, we have assumed that lexical diversity and density correlate directly with translation quality; i.e. the more diverse and dense a translation the better. In this regard, we acknowledge that in translation there is a tension between diversity and consistency, especially in technical translation. At the same time, none of our datasets falls under the domain of technical texts.

We also acknowledge that our study is based on rather superficial linguistic features, either at surface or morphological level (PoS tags). For future work, therefore, we plan to explore the use of additional features, especially relying on deeper linguistic analyses. In addition, we plan to study the overlap between multiple HTs and PEs for the same text, to assess whether it is higher between PEs, which would indicate a higher degree of homogeneisation.

Another line we would like to pursue is that of automatic discrimination between PE and HT. While this has been shown to be possible with a high degree of accuracy between original texts and HTs, this is not the case for PE versus HT in the attempts conducted to date (Daems et al., 2017).

Finally, we would like to point out that all our code and data are publicly released,[15] so we encourage interested parties to use these resources to conduct further analyses.

Acknowledgments

I would like to thank Luisa Bentivogli, Joke Daems, Michael Farrell, Lieve Macken and Maja Popović for pointing me to relevant datasets and related papers and for insightful discussions on the topic of this paper. I would also like to thank the reviewers; their comments have definitely led to improve this paper.

References

Avramidis, Eleftherios, Aljoscha Burchardt, Sabine Hunsicker, Maja Popovic, Cindy Tscherwinka, David Vilar, and Hans Uszkoreit. 2014. The taraxü corpus of human-annotated machine translations. In *LREC*, pages 2679–2682.

Baker, Mona. 1993. Corpus linguistics and translation studies: Implications and applications. *Text and technology: In honour of John Sinclair*, 233:250.

Bentivogli, Luisa, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. 2016. Neural versus phrase-based machine translation quality: a case study. In Su, Jian, Xavier Carreras, and Kevin Duh, editors, *Proceedings of the 2016 Conference on Empirical*

[15] `https://github.com/antot/posteditese_mtsummit19`

Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016, pages 257–267. The Association for Computational Linguistics.

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.

Bowker, Lynne. 2009. Can machine translation meet the needs of official language minority communities in canada? a recipient evaluation. *Linguistica Antverpiensia, New Series–Themes in Translation Studies*, (8).

Cettolo, Mauro, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, Roldano Cattoni, and Marcello Federico. 2015. The iwslt 2015 evaluation campaign. In *IWSLT 2015, International Workshop on Spoken Language Translation*.

Daems, Joke, Orphée De Clercq, and Lieve Macken. 2017. Translationese and post-editese : how comparable is comparable quality? *LINGUISTICA ANTVERPIENSIA NEW SERIES-THEMES IN TRANSLATION STUDIES*, 16:89–103.

Farrell, Michael. 2018. Machine Translation Markers in Post-Edited Machine Translation Output. In *Proceedings of the 40th Conference Translating and the Computer*, pages 50–59.

Fiederer, Rebecca and Sharon O'Brien. 2009. Quality and machine translation: A realistic objective. *The Journal of Specialised Translation*, 11:52–74.

Garcia, Ignacio. 2010. Is machine translation ready yet? *Target. International Journal of Translation Studies*, 22(1):7–21.

Green, Spence, Jeffrey Heer, and Christopher D Manning. 2013. The efficacy of human post-editing for language translation. *Chi 2013*, pages 439–448.

Guerberof, Ana. 2009. Productivity and quality in mt post-editing. In *MT Summit XII-Workshop: Beyond Translation Memories: New Tools for Translators MT*. Citeseer.

Hassan, Hany, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, Will 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. `https://arxiv.org/abs/1803.05567`.

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

Mauro, Cettolo, Niehues Jan, Stüker Sebastian, Bentivogli Luisa, Cattoni Roldano, and Federico Marcello. 2016. The iwslt 2016 evaluation campaign. In *International Workshop on Spoken Language Translation*.

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.

Scarpa, Federica. 2006. Corpus-based quality-assessment of specialist translation: A study using parallel and comparable corpora in english and italian. *Insights into specialized translation–linguistics insights. Bern: Peter Lang*, pages 155–172.

Stolcke, Andreas. 2002. Srilm-an extensible language modeling toolkit. In *Seventh international conference on spoken language processing*.

Straka, Milan and Jana Straková. 2017. Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada, August. Association for Computational Linguistics.

Straka, Milan, Jan Hajic, and Jana Straková. 2016. Udpipe: Trainable pipeline for processing conll-u files performing tokenization, morphological analysis, pos tagging and parsing. In *LREC*.

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: Volume 1, Long Papers*, pages 1063–1073, Valencia, Spain, April. Association for Computational Linguistics.

Toral, Antonio, Sheila Castilho, Ke Hu, and Andy Way. 2018. Attaining the unattainable? reassessing claims of human parity in neural machine translation. In *Proceedings of the Third Conference on Machine Translation, Volume 1: Research Papers*, pages 113–123, Belgium, Brussels, October. Association for Computational Linguistics.

Toury, Gideon. 2012. *Descriptive translation studies and beyond: Revised edition*, volume 100. John Benjamins Publishing.

Čulo, Oliver and Jean Nitzke. 2016. Patterns of terminological variation in post-editing and of cognate use in machine translation in contrast to human translation. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation, EAMT 2017, Riga, Latvia, May 30 - June 1, 2016*, pages 106–114. European Association for Machine Translation.