Post-editors as Gatekeepers of Lexical and Syntactic Diversity: Comparative Analysis of Human Translation and Post-editing in Professional Settings

Lise Volkart
FTI/TIM, University of Geneva
Switzerland
lise.volkart@unige.ch

Pierrette Bouillon
FTI/TIM, University of Geneva
Switzerland
pierrette.bouillon@unige.ch

Abstract

This paper presents a comparative analysis between human translation (HT) and post-edited machine translation (PEMT) from a lexical and syntactic perspective to verify whether the tendency of neural machine translation (NMT) systems to produce lexically and syntactically poorer translations shines through after post-editing (PE). The analysis focuses on three datasets collected in professional contexts containing translations from English into French and German into French. Through a comparison of word translation entropy (HTRa) scores, we observe a lower degree of lexical diversity in PEMT compared to HT. Additionally, metrics of syntactic equivalence indicate that PEMT is more likely to mirror the syntactic structure of the source text in contrast to HT. By incorporating raw machine translation (MT) output into our analysis, we underline the important role post-editors play in adding lexical and syntactic diversity to MT output. Our findings provide relevant input for MT users and decision-makers in language services as well as for MT and PE trainers and advisers.

1 Introduction

Post-editing (PE) has now largely proved to be a good alternative to purely human translation (HT) in professional contexts. By allowing certain productivity gains without negatively affecting the quality of the final translation (Daems, 2016; Läubli et al., 2019), MT and PE have found their place in professional translation workflows. Nevertheless, translators often express mixed feelings towards MT. On one hand, the tool is appreciated for its help when dealing with high workloads and time constraints, but on the other hand, it is perceived as a threat to translation's creativity, originality and naturalness (Alvarez-Vidal et al., 2020; Girletti, 2024). These concerns are legitimate: numerous studies have revealed the NMT tendency to produce an output that is less lexically varied and syntactically closer to source text than HT (Vanmassenhove et al., 2019; Toral, 2019; Vanmassenhove et al., 2021; Webster et al., 2020; Ahrenberg, 2017; Shaitarova et al., 2023; Luo et al., 2024). Furthermore, some studies have found measurable differences on parallel corpora of HT and post-edited machine translation (PEMT) in terms of lexical diversity, lexical density and sentence length, among others, suggesting the existence of a *post-editese* phenomenon (Castilho et al., 2019; Castilho and Resende, 2022; Toral, 2019). However, Volkart and Bouillon (2023), demonstrated the difficulty of generalising these findings over different corpora, domains and language pairs, particularly when analysing authentic comparable corpora. Such corpora, in which HT and PEMT are the translations of different source texts, necessitate analysis with metrics that encompass the attributes of the source. Although demanding, the study of authentic HT and PEMT corpora is crucial for developing a detailed understanding of the distinct characteristics of PEMT output in professional contexts.

In this study, we compare the lexical and syntactic characteristics of authentic HT and PEMT output produced in professional contexts relying

Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1), pages 387–395
June 24-27, 2024 ©2024 European Association for Machine Translation

on metrics suitable for the study of comparable corpora. Whenever possible, raw MT output is added to the analysis to confirm initial premises on raw MT, as well as findings by previous studies. We measure the variety of translation solutions by automatically computing average word translation entropy (HTra) (Carl et al., 2016) and the syntactic equivalence between source and target based on three syntactic equivalence metrics from the ASTReD library (Vanroy et al., 2021) to answer the following research question:

Is the NMT tendency toward lexical impoverishment and source sentence structure mirroring discernible in the PEMT final product?

Throughout our analysis, we make the following observations:

1. PEMT final output is affected by NMT bias in terms of lexical diversity and presents lower levels of translation variety

2. Syntactic shining through from raw NMT occurs in PEMT output but remains limited

3. PE adds significant levels of lexical and syntactic variety to MT output

By including three authentic corpora, two language pairs, various state-of-the-art NMT systems and carefully selected metrics, our study contributes to improve our understanding of the impact of MT integration on translated language. Our findings provide valuable insights to inform decisions about where and when to use or not to use MT and can contribute to enhancing PE training programs and refining best practices. Finally, the comparison between HT, PEMT and raw MT reaffirms the essential role played by human post-editors.

Section 2 gives a brief overview of previous relevant studies. Datasets and corpora are described is Section 3 and experimental setup in Section 4. We present and comment our results in Section 5. Section 6 briefly presents illustrative examples. Finally, Section 7 reports our conclusions and main findings.

2 Related work

Lexical level Vanmassenhove et al. (2019), compared statistical machine translation (SMT), NMT and HT in terms of lexical richness, to verify the hypothesis according to which data-driven MT systems, due to their probabilistic nature, would tend to favour more frequent words and disregard less frequent ones and therefore produce a lexically less diverse output than HT. Their experiment, conducted on 12 different MT systems (SMT and NMT with different architectures and with and without using backtranslated data) and two languages directions (EN-FR and EN-ES), confirmed this hypothesis. The three investigated lexical richness metrics (Yule's I, type/token ratio and measure of textual lexical diversity) indicated a lower lexical richness in the MT output in comparison to HT. With a further analysis on word frequencies, the authors demonstrated the tendency of MT systems to increase the frequency of already frequent words while decreasing the frequency of less frequent ones. This tendency toward overgeneralisation was again observed by Vanmassenhove et al. (2021) when measuring the difference in lexical and morphological richness between an MT system's training data and its output for the language directions EN-FR and EN-ES. They measured a loss of lexical and morphological richness between the training data and the system's output. Webster and al. (2020) also observed a loss of lexical richness and a homogenisation of lexicon with NMT when comparing lexical richness of literary excerpts translated by humans and by two online NMT systems. Although pursuing a slightly different goal (i.e. comparing raw MT, PEMT and revised PEMT), the work by Macken et al. (2022) is worth mentioning here, particularly because the authors relied, among others, on the average automatic word translation entropy (denoted AWTE in their paper) to assess lexical richness of the different translation modes. Overall, their experiment showed that PE and revision tend to increase the lexical variety of the raw MT, with AWTE being the most unequivocal of the three metrics used (AWTE, TTR and Mass Index).

Syntactic level As for the syntactic profile of MT outputs, several studies investigated the syntactic similarity between source and target for HT and PEMT. In 2017, Ahrenberg (2017) found out that, when translating from English into Swedish, NMT tends to produce an output that mimic the source structure, performing less word re-ordering than human translators. Comparing HT and generic NMT on literary excerpts with the help of wordcross and ASTrED metrics (Vanroy et al., 2021), Webster at al. (2020) found out NMT tends to re-

main syntactically closer to the source structure. The same tendency was observed by Shaitarova et al. (2023), who tested syntactic equivalence between source and target using the ASTrED tool to compute cross-alignments on several large corpora. Their comparison of HT and NMT from different commercial systems indicated a general tendency of NMT systems to reproduce the syntax of the source, whereas HT appears to be more creative on this aspect. It is worth noting that in this experiment, out of the 4 tested NMT systems, DeepL appeared as the one producing the most syntactically diversified output. Finally, in an extensive study comparing NMT and HT in terms of morphosyntactic divergence between source and target on three language directions, Luo et al. (2024) found out that NMT tends to produce less diverse morphosyntactic patterns and more one-to-one alignments than HT.

3 Datasets

Our experiment is based on three authentic datasets containing professional translations collected from in-house language services[1]. Each dataset contains a balanced amount of HT and PEMT segments with their respective source. Dataset ENfr1 was compiled from the same data as in Volkart and Bouillon (2022) and contains translations from English into French extracted from documents of the European Investment Bank (EIB). Dataset ENfr2 and dataset DEfr are derived from the dataset described in Volkart and Bouillon (2023). Dataset ENfr2 contains translations from English into French shared with us by a sports organisation based in Switzerland, while dataset DEfr contains translations from German into French collected from an insurance company. For all datasets, raw MT used for PE came from various state-of-the-art NMT systems (generic and/or customised). Table 1 presents the size of each dataset and corpus. In addition to PEMT data, we added raw MT of the PEMT source data to ENfr1 and ENfr2 datasets. This raw MT was generated for this experiment using DeepL Pro[2] [3]. Original raw MT is not saved by the language services during the PE process, which restricts our analysis of authentic data to PEMT product. However, we deem informative to include an example of raw MT output, although artificial, in our analysis. It allows us, among others, to verify if the tendencies observed by previous studies on the lexical and syntactic profile of MT outputs are indeed to be seen in our data.

[1]All services shared their data on a voluntary basis. Agreements between researchers and organisations were signed when needed and data was anonymized when required.

[2]in february 2024

[3]Data provider of the DEfr corpus did not allow us to translate their corpus using an online MT system

4 Experiment

4.1 Variety of translation solutions

To measure the variety of translation solutions, we rely on the Word translation entropy metric (denoted HTra) (Carl et al., 2016). HTra is computed as the sum over all observed word translation probabilities $p(s \rightarrow t_i)$ of a given source text word s into target text word $t_i...n$ multiplied with their information content $I(p) = -log_2(p)$ (Carl et al., 2016) as is the following equation:

$$HTra(s) = -\Sigma_{i=1}^{n} p(s \rightarrow t_i) \times log_2(p(s \rightarrow t_i))$$

This score reflects, for a given source word, the amount of translation alternatives and their distribution in the target (Bangalore et al., 2016; Gilbert et al., 2023). The higher the HTra, the higher the variety in the translation of that source word in the target corpus. Compared to the TTR-based scores (such as TTR (Scott, 2019), STTR (Scott, 2019), MSTTR (Malvern and Richards, 2002) or MATTR (Covington and McFall, 2010)) often used to compare lexical richness of HT and MT/PEMT, HTra offers two main advantages: first, it is computed on the target corpus given its source and therefore allows us to compare translations from different source texts more easily (whereas TTR requires us to take into account the influence of the source while comparing the target, see Volkart and Bouillon (2022) and Volkart and Bouillon (2023) for a more detailed discussion on this aspect), and second, it encompasses two different aspects of lexical/translation richness that are the number of unique translation solutions, and also their distribution (does one solution account for 90% of the occurrences or are all translation solutions equally used by the translator?). Then, in addition to being more appropriate regarding our corpus design, HTra captures more information on the lexical richness of different translations than TTR-based metrics.

Computing HTra for a given source word requires the extraction all occurrences of that source

Dataset	Trans. mode	# segment pairs	# source tokens
ENfr1	HT	1,852	40,560
	PEMT	1,852	41,803
ENfr2	HT	2,280	43,379
	PEMT	2,280	49,896
DEfr1	HT	7,769	106,864
	PEMT	7,769	106,673

Table 1: Number of segments and source tokens for each dataset and translation mode

word with their respective translations in the target corpus. Whereas it can be done manually on small corpora or for a selection of source words such as in Volkart and Bouillon (2022), where it was computed for a set of 20 adverbs, it can rapidly become impossible to apply on large corpora.

We computed HTra automatically using ad hoc python scripts. Word alignment was performed with awesome-align (Dou and Neubig, 2021), a neural word aligner based on multilingual BERT, without fine-tuning. Out of this automatic alignment, we extracted the list of source-target pairs for content words (adverbs, adjectives, nouns and verbs) and grouped source words aligned with multiple target words together to form one-to-many alignments. Non-aligned source words were added to the list as non-translated. Tagging and lemmatization were performed in parallel using SpaCy's transformer models for English, French and German [4]. We computed HTra for all content source lemmas that occur at least three times in both HT and PEMT source corpora.

This automatic HTra computation pipeline was validated against manually computed scores from Volkart and Bouillon, on the same corpus and the same subset of adverbs (2022). Pearson's correlation coefficients between automatic and manual HTra scores are respectively 0,83 for HT and 0,81 for PEMT. These high levels of correlation validate the automatic calculation method as well as the quality of the automatic word alignment.

4.2 Syntactic equivalence

To measure the impact of PEMT on the syntactic level, we used the ASTrED python library (Vanroy et al., 2021) to compute three metrics of syntactic equivalence, namely the label changes, the Syntactically Aware Cross (SACr) and the Aligned syntactic tree edit distance (ASTrED). Those metrics aim at capturing syntactic equivalence between a source and a target segment based on differences in word/word group order, differences in dependency labels and differences in syntactic structures (Vanroy et al., 2021). The ASTrED library relies on Stanza parser (Qi et al., 2020) for universal dependency parsing and an adapted version of awesome-align (Dou and Neubig, 2021) for word alignment (Vanroy et al., 2021).

Label changes Label changes correspond, for a given source-target sentence pair, to the number of source-target word-aligned pairs that have different dependency labels, normalised by the total number of alignments for that sentence. This metric captures the linguistic differences between aligned words on the surface level (Vanroy et al., 2021).

SACr SACr quantifies the degree of reordering of word sequences that occurred between source and target (Vanroy et al., 2021). Words are grouped together according to their relation in the dependency tree to form linguistically motivated word sequences. Source and target word sequences are then aligned based on word alignments and SACr value is computed by dividing the number of cross-alignments normalised by the total number of alignments. SACr captures the surface word order differences between the source and target sentences.

ASTrED ASTrED captures the source and target structural differences on a deeper level by comparing dependency trees while taking word alignments into account. The computed tree edit distance is normalised by the average number of source and target words (Vanroy et al., 2021).
For further details and illustrated examples on these metrics, we invite the reader to refer to Vanroy et al. (2021).

[4] https://spacy.io/models

5 Results

5.1 Lexical richness

HTra was automatically computed on content lemmas for HT, PEMT and raw MT[5]. Average scores by POS categories and for all content lemmas for HT, PEMT and raw MT corpora are presented in Table 2.

All POS categories together, the first thing we observe is that the raw MT generated with DeepL for this experiment presents a much lower variety of translation solutions compared to HT. For all categories together, the HTra score is more than 20% lower for raw MT. This confirms what has been observed in previous studies regarding the tendency of MT systems to narrow the range of translation solutions by increasing the frequency of already frequent words while decreasing the frequency of less frequent ones. This tendency shines through in the PEMT output which exhibits a generally lower HTra score compared to HT for all three datasets. Datasets ENfr2 and DEfr present very similar results, with HTra almost 8% lower for PEMT in contrast to HT. This loss of translation solution variety is less marked in the ENfr1 dataset, but still to be seen.

Those scores indicate that post-editors presumably add significant amount of lexical variety to the MT output, but still not enough to reach the level of variation from HT.

Looking at HTra scores by POS category separately, we see that the observed loss of translation variety is spread differently across categories for the different datasets. For ENfr1, adverbs show the biggest loss of variety in PEMT, while this loss is very limited for verbs. For ENfr2, on the contrary, the loss of variety affects primarily nouns and verbs, while for adverbs we even observe a higher translation variety in PEMT. Interestingly, the loss of translation solution variety in raw MT seems to correlate with the loss of translation solution variety for PEMT for this dataset. Finally, in DEfr, the loss of variety in PEMT is more evenly spread across categories, with a slightly stronger effect for adjectives and adverbs. Here, different POS categories appear to be differently affected by the loss of translation variety in PEMT depending on the dataset and, presumably, on the language pair.

[5]To prevent the results from being overly biased by non-frequent or topic-related lemmas, HTra was computed for content lemmas occurring at least three time in both source corpora

5.2 Syntactic equivalence

Table 3 presents the average scores for HT for all three metrics and the relative differences for PEMT and raw MT when available. For all three metrics, a lower score indicates a higher level of syntactic equivalence between source and target. Similarly to what we observe for HTra, raw MT differs significantly from HT for both English into French datasets, with all three metrics indicating that the target tends to be syntactically closer to the source for raw MT. Once again, it is coherent with what could be observed in other studies. The difference between HT and PEMT is less straightforward. For the ENfr1 dataset, word sequence reordering, as measured by SACr, and label changes are more frequent in PEMT, whereas tree edit distance is slightly lower for PEMT. As for ENfr2, metrics show that PEMT is syntactically closer to source than HT, with significantly less label changes and lower tree edit distance. As for DEfr, SACr and label changes are not significantly different for PEMT and HT, but ASTrED points toward more similarity between source and target on the deeper level in PEMT. These results show that PE clearly blurs the line between HT and MT on the syntactic level. Post-editors play a major role in adding syntactic variety to the MT output during post-editing especially on the surface level by adding large amounts of word reordering and dependency label change. On a deeper level however, PEMT stays closer to the source than HT as expressed by consistently lower ASTrED scores.

6 Examples

Loss of translation solution variety: to illustrate what a lower HTra score concretely means, we present examples showing the translation solutions distribution for particular lemmas in Figures 1 and 2 in Appendix A. Figure 1 shows the distribution of translation solutions for the noun “impact” in HT, PEMT and raw MT within the ENfr2 dataset. While PEMT and HT exhibit an equal number of translation solutions, their frequencies are more evenly dispersed in HT, resulting in a HTra score of 2.84 for HT compared to 2.79 for PEMT. In contrast, raw MT yields only two distinct translation solutions (with the absence of translation considered as a translation “choice”), where one solution overwhelmingly dominates,

	ENfr1			ENfr2			DEfr	
Corpus	**HT**	**PEMT**	**RawMT**	**HT**	**PEMT**	**RawMT**	**HT**	**PEMT**
ADJ	1.32	-1.52%	-17.42%*	1.45	-2.76%	-20.00%$^{\diamond}$	1.69	-9.47%$^{\diamond}$
ADV	1.51	-4.64%	-15.89%	1.69	+2.37%	-10.65%	1.93	-9.33%$^{\diamond}$
NOUN	1.07	-2.80%	-24.30%$^{\diamond}$	1.26	-10.32%$^{\diamond}$	-33.33%$^{\diamond}$	1.16	-6.90%$^{\diamond}$
VERB	1.92	-0.52%	-23.44%$^{\diamond}$	2.16	-9.72%$^{\diamond}$	-23.15%$^{\diamond}$	1.90	-7.89%$^{\diamond}$
All	**1.34**	**-1.49%**	**-21.64%**$^{\diamond}$	**1.54**	**-7.79%**$^{\diamond}$	**-25.97%**$^{\diamond}$	**1.53**	**-7.84%**$^{\diamond}$

Table 2: Average HTra scores for HT and relative difference for PEMT and raw MT for all content lemmas and each POS category. *indicate significance at $p < 0.005$ and $^{\diamond}$ at $p < 0.001$ Significance was tested using Mann Whitney non-parametric test.

	ENfr1			ENfr2			DEfr	
Corpus	**HT**	**PEMT**	**RawMT**	**HT**	**PEMT**	**RawMT**	**HT**	**PEMT**
ASTrED	0.6153	-1.67%	-3.79%$^{\diamond}$	0.6348	-7.29%$^{\diamond}$	-14.07%$^{\diamond}$	0.6518	-3.01%$^{\diamond}$
SACr	0.2702	+6.62%$^{\diamond}$	-24.46%*	0.2558	-3.36%	-40.89%$^{\diamond}$	0.3185	-0.78%
Label Ch.	0.1982	+1.11%	-5.90%$^{\diamond}$	0.2186	-7.55%$^{\diamond}$	-11.89%$^{\diamond}$	0.2292	+0.26%

Table 3: Syntactic equivalence scores for HT and relative difference for PEMT and raw MT. *indicate significance at $p < 0.005$ and $^{\diamond}$ at $p < 0.001$. Significance was tested using Mann Whitney non-parametric test

resulting in an HTra score of 0.24. Looking at the adverb "also" within the ENfr1 dataset presented in Figure 2, we note that in this situation the high HTra score for HT (1.27) is principally due to the number of different translation solutions, more than to their frequency distribution. PEMT achieves an HTra score of only 1.02 while raw MT, due to the strong dominance of the most frequent solution barely reaches 0.85. These examples show how the loss of lexical diversity occurs in PEMT through the loss of translation solution variety and how the use of MT can lead to a reinforcement of the most frequent translation solutions at the expense of the less frequent ones.

Syntactic equivalence: Figures 3 and 4 show the word alignments between source and target for two sentences extracted from our datasets. They illustrate two contrasting examples regarding syntactic equivalence between source and target. In Figure 3, the target sentence presents a high level of syntactic equivalence according to the three computed scores (ASTrED = 0.29, SACr = 0.05, Label change = 0.16) and this is intuitively expressed in the word alignment. The target sentence is an almost one-to-one translation of the source with minimal word reordering and dependency label changes. Figures 4 in contrast presents the word alignments with higher scores and therefore less syntactic equivalence (ASTrED = 0.78, SACr = 0.58, Label change = 0.25). Here again, just by looking at the word alignment, it is clear that the target presents higher levels of word reordering and structural differences compared to the source sentence.

7 Conclusion

This paper compares authentic sets of HT and PEMT produced in three different professional contexts for the language directions English into French and German into French, with additional analysis incorporating raw MT output from DeepL for two of the datasets. The objective is to compare HT and PEMT in terms of lexical and syntactic variety to verify whether the general tendency of NMT systems to produce lexically and syntactically less varied output still shines through after a PE step performed by professional translators. Using HTra for the lexical aspects and ASTrED, SACr and dependency label changes (Vanroy et al., 2021) for the syntactic aspects, we note the strong tendency of raw NMT (in this case DeepL) to produce lexically less varied translations that tend to mirror the source sentence structure. This tendency is strongly attenuated by the PE step, with PEMT output being generally closer to HT than to raw MT on both aspects. This indicates that post-editors presumably add significant levels of lexical and syntactic variety to the MT output ("presumably", because raw MT under analysis is not the one originally used for PEMT, but we assume it reflects the general level of lexical and syntactic variation of NMT systems). Still, the final PEMT output does not systematically reach the same level

of variety as HT, especially on the lexical level. For all datasets PEMT exhibits lower levels of translation solution variety. A tendency towards more syntactic equivalence between source and target in PEMT is clear for one dataset but more nuanced for the two others. These findings are particularly relevant in contexts where lexical and syntactic variety are regarded as criteria for assessing translation quality.

Furthermore, our work highlights the crucial importance of PE, not only in ensuring the accuracy of the target text, but also in maintaining an adequate level of lexical diversity and syntactic naturalness in the final translation. While this aspect may seem unimportant for certain types of texts, it holds significant relevance for others. In many cases, (human) translation is not only about overcoming language barriers but also about producing "texts that satisfy the linguistic norms of a target culture and are adapted to the assumed knowledge of its reader" (Ahrenberg, 2017, 1). It is also of utmost importance considering the fact that PEMT output is likely to be re-used to train NMT systems, and therefore to amplify over and over the already existing biases.

Finally, we emphasise the relevance of our findings for the improvement of post-editing training programs and guidelines. While translators are still today often advised to stick to the TAUS PE guidelines (TAUS and CNGL, 2010) and to not intervene on the stylistic level, we are convinced that adding lexical and syntactic diversity (even when not strictly necessary from micro-level perspective) to MT output is essential to preserve the quality of translated text at the macro-level.

Acknowledgement

We would like to thank the language services who kindly accepted to share their translation memories for this research project. We would also like to thank the reviewers for their insightful comments and feedback.

References

Ahrenberg, Lars. 2017. Comparing machine translation and human translation: A case study. In *RANLP 2017: The First Workshop on Human-Informed Translation and Interpreting Technology (HiT-IT)*, pages 21–28. Association for Computational Linguistics.

Alvarez-Vidal, Sergi, Antoni Oliver, and Toni Badia. 2020. Post-editing for professional translators: cheer or fear? *Tradumàtica*, (18):0049–69.

Bangalore, Srinivas, Bergljot Behrens, Michael Carl, Maheshwar Ghankot, Arndt Heilmann, Jean Nitzke, Moritz Schaeffer, and Annegret Sturm. 2016. Syntactic variance and priming effects in translation. In *New directions in empirical translation process research*, pages 211–238. Springer.

Carl, Michael, Moritz Schaeffer, and Srinivas Bangalore. 2016. The CRITT translation process research database. In *New directions in empirical translation process research*, pages 13–54. Springer.

Castilho, Sheila and Natália Resende. 2022. Post-editese in literary translations. *Information*, 13(2):66. Publisher: Multidisciplinary Digital Publishing Institute.

Castilho, Sheila, Natália Resende, and Ruslan Mitkov. 2019. What influences the features of post-editese? a preliminary study. In *Proceedings of the Human-Informed Translation and Interpreting Technology Workshop (HiT-IT 2019)*, pages 19–27. Varna, Bulgaria.

Covington, Michael A. and Joe D. McFall. 2010. Cutting the Gordian Knot: The Moving-Average Type–Token Ratio (MATTR). *Journal of Quantitative Linguistics*, 17(2):94–100.

Daems, Joke. 2016. *A translation robot for each translator?: A comparative study of manual translation and post-editing of machine translations: Process, quality and translator attitude*. PhD Thesis, Ghent University.

Dou, Zi-Yi and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Gilbert, Devin, Cristina Toledo-Báez, Michael Carl, and Haydeé Espino. 2023. Impact of word alignment on word translation entropy and other metrics. In Lacruz, Isabel, editor, *Translation in Transition: Human and machine intelligence*, page 203. Publisher: John Benjamins Publishing Company.

Girletti, Sabrina. 2024. *Working with Pre-translated Texts: Investigating Machine Translation Post-editing and Human Translation Revision at Swiss Corporate In-house Language Services*. Ph.D. thesis, University of Geneva.

Läubli, Samuel, Chantal Amrhein, Patrick Düggelin, Beatriz Gonzalez, Alena Zwahlen, and Martin Volk. 2019. Post-editing productivity with neural machine translation: An empirical assessment of speed and quality in the banking and finance domain. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 267–272, Dublin, Ireland. European Association for Machine Translation.

Luo, Jiaming, Colin Cherry, and George Foster. 2024. To Diverge or Not to Diverge: A Morphosyntactic Perspective on Machine Translation vs Human Translation. *arXiv preprint arXiv:2401.01419*.

Macken, Lieve, Bram Vanroy, Luca Desmet, and Arda Tezcan. 2022. Literary translation as a three-stage process: machine translation, post-editing and revision. In *23rd Annual Conference of the European Association for Machine Translation*, pages 101–110. European Association for Machine Translation.

Malvern, David and Brian Richards. 2002. Investigating accommodation in language proficiency interviews using a new measure of lexical diversity. *Language Testing*, 19(1).

Qi, Peng, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Scott, Mike. 2019. WordSmith tools manual.

Shaitarova, Anastassia, Anne Göhring, and Martin Volk. 2023. Machine vs. Human: Exploring Syntax and Lexicon in German Translations, with a Spotlight on Anglicisms. In *The 24rd Nordic Conference on Computational Linguistics*.

TAUS and CNGL. 2010. Machine Translation Post-Editing Guidelines.

Toral, Antonio. 2019. Post-editese: an exacerbated translationese. In *Proceedings of MT Summit XVII*, volume 1, pages 273 – 281. Dublin, Ireland.

Vanmassenhove, Eva, Dimitar Shterionov, and Andy Way. 2019. Lost in translation: Loss and decay of linguistic richness in machine translation. In *Proceedings of MT Summit XVII*, volume 1, pages 222 – 232. Dublin, Ireland.

Vanmassenhove, Eva, Dimitar Shterionov, and Matthew Gwilliam. 2021. Machine Translationese: Effects of Algorithmic Bias on Linguistic Complexity in Machine Translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2203–2213, Online, April. Association for Computational Linguistics.

Vanroy, Bram, Orphée De Clercq, Arda Tezcan, Joke Daems, and Lieve Macken. 2021. Metrics of syntactic equivalence to assess translation difficulty. In *Explorations in empirical translation process research*, pages 259–294. Springer.

Volkart, Lise and Pierrette Bouillon. 2022. Studying Post-Editese in a Professional Context: A Pilot Study. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 71–79.

Volkart, Lise and Pierrette Bouillon. 2023. Are post-editese features really universal? In Orăsan, Constantin, Ruslan Mitkov, Gloria Corpas Pastor, and Johanna Monti, editors, *Proceedings of the International Conference on Human-Informed Translation and Interpreting Technology (HiT-IT 2023)*, pages 294–304, Naples.

Webster, Rebecca, Margot Fonteyne, Arda Tezcan, Lieve Macken, and Joke Daems. 2020. Gutenberg goes neural: Comparing features of dutch human translations with raw neural machine translation outputs in a corpus of english literary classics. In *Informatics*, volume 7, page 32. MDPI. Issue: 3.

Appendix A. Examples

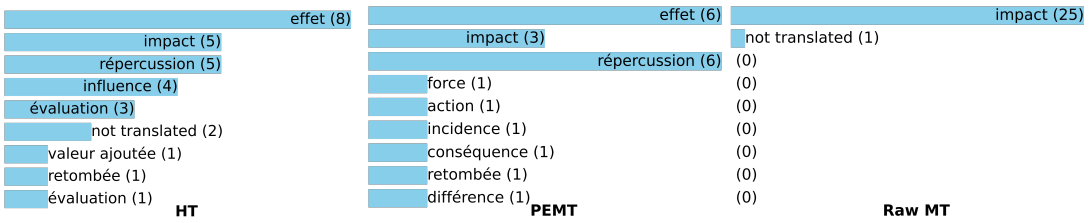

Figure 1: Translation solutions distribution for the noun "impact" in ENfr2 HT, PEMT and raw MT.

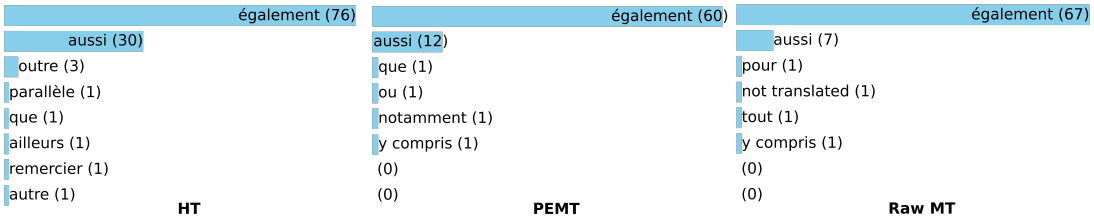

Figure 2: Translation solution distribution for the adverb "also" in ENfr1 HT, PEMT and raw MT.

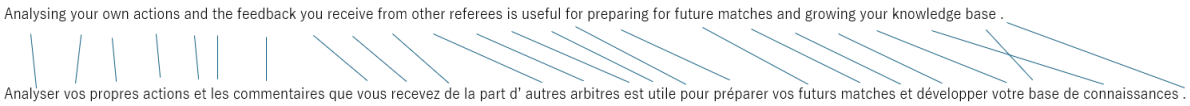

Figure 3: Example of a source-target sentence pair presenting high levels of syntactic equivalence, with automatic word alignments.

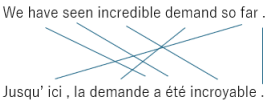

Figure 4: Example of a source-target sentence pair presenting low levels of syntactic equivalence, with automatic word alignments.