On the differences between human translations

Maja Popović
ADAPT Centre
School of Computing
Dublin City University, Ireland
maja.popovic@adaptcentre.ie

Abstract

Many studies have confirmed that translated texts exhibit different features than texts originally written in the given language. This work explores texts translated by different translators taking into account expertise and native language. A set of computational analyses was conducted on three language pairs, English-Croatian, German-French and English-Finnish, and the results show that each of the factors has certain influence on the features of the translated texts, especially on sentence length and lexical richness. The results also indicate that for translations used for machine translation evaluation, it is important to specify these factors, especially when comparing machine translation quality with human translation quality.

1 Introduction

Many studies have demonstrated that translated texts (human translations, HTs) have different lexical, syntactic and other textual features than texts originally written in the given language (originals). These special traits of HTs are result of a compromise between two often antagonised aspects of the translation process: fidelity to the source text and naturalness of the generated target language text. Although all studies confirm the existence of unique HT features, two categories of these features are distinguished in the literature. One category, "translation universals", represents a general set of features shared by all translations, independent of the characteristics of involved languages (Baker et al., 1993). Another category, "interference", reflects the impact of the source language, the "trace" which the source language leaves in the translation (Toury, 1979). Some studies investigate and demonstrate the existence of both categories, sometimes called "source universals" and "target universals" (Chesterman, 2004; Koppel and Ordan, 2011).

Our research aims to find out whether differences between translators have any influence on the text features. We investigate impact of the translator's expertise and native language. We present results of a computational analysis of a set of HTs originating from the news domain and involving three distinct language pairs, English-Croatian, German-French and English-Finnish. The analysis is guided by the following research questions:

RQ1 Are there differences between HTs related to translator's expertise?

RQ2 Are there differences between HTs related to translator's native language and translation direction? (from or into translator's native language)

The main contribution of this work is empirical, showing evidence of differences between text features of HTs produced by different translators. We expect our findings to motivate and drive future research in this direction in order to better understand these differences by identifying and analysing underlying linguistic phenomena.

Moreover, differences between HTs may have practical impact on evaluation of machine translation (MT) systems. Several recent studies (Toral et al., 2018; Läubli et al., 2018; Zhang and Toral, 2019; Freitag et al., 2019) have shown that the

translation direction has impact on the results of evaluation of MT outputs, so that it is important to specify whether originals or HTs were used as source texts for MT systems. Taking into account these studies and the findings reported in this work, potential effects of translators' backgrounds on MT should be investigated too.

2 Related work

Analysis of translated texts A lot of work has been done exploring differences between HTs and originals. Some studies (Baker et al., 1993) have emphasised the existence of "translation universals", general features of translated texts, "simplification" and "explicitation" being the most well-known. Other studies (Toury, 1979) have pointed out the influence of the source language, "interference", whereas some (Chesterman, 2004) concentrate on both categories, called "S-universals" and "T-universals".

Since many text features can be measured quantitatively, a number of publications demonstrated that HTs can be automatically distinguished from originals (Baroni and Bernardini, 2006; Koppel and Ordan, 2011; Volansky et al., 2015; Rabinovich and Wintner, 2015; Rubino et al., 2016). The features used for the classifiers are partly motivated by the theoretical categories mentioned above, however many features are not directly related to a particular category, and many can belong to more than one category. The most common features are lexical variety (percentage of distinct words in a text), lexical density (sometimes called information density, percentage of content words in a text), sentence length, word length, as well as frequencies of certain POS categories, function words and collocations.

Rabinovich et al. (2016) include analysis of non-native texts, namely texts originally written in the given language but by non-native speakers. They found that these texts generally exhibit different features than native originals and HTs, thus representing yet another text category. On the other hand, their features are closer to those of HTs than to native originals, indicating the influence ("interference") of the native language.

In addition to analysis of HTs, more and more publications report analysis of machine translated texts. Ahrenberg (2017) compares MT outputs with HTs by means of automatically calculated text features as well as by manual analysis of divergences (shifts) from the source text. The main finding is that MT output is much more similar to the source text than HT. Another study of machine translated texts (Vanmassenhove et al., 2019) reports significantly lower lexical richness in MT outputs in comparison to originals and HTs.

Post-editing (PE) of MT outputs has lead to yet another type of translated text which has been analysed extensively in the recent years (Čulo and Nitzke, 2016; Daems et al., 2017; Farrell, 2018; Toral, 2019; Castilho et al., 2019). These studies demonstrated that PEs represent an additional text category with the features lying between those of HTs and of MT outputs.

Relations between machine and human translation As machine translation (MT) technology improves, more and more work has been done on investigating relations between different aspects of MT and HT direction. First publications on this topic (Kurokawa et al., 2009; Lembersky et al., 2013) demonstrated that the direction of HT plays an important role for building a statistical MT system, and recommend training on parallel corpora which were translated in the same direction as the MT system (i.e. using originals as source and HTs as target).

Recently, several publications (Läubli et al., 2018; Toral et al., 2018; Freitag et al., 2019; Zhang and Toral, 2019) demonstrated that the translation direction plays an important role both for human as well as for automatic evaluation of MT systems. Before these findings were published, this aspect has not been taken into account at all in the MT community.[1] Afterwards, as a consequence, using only originals as source test texts and HTs as reference test texts has become a common practice in the WMT shared tasks[2] from 2019. The main reason is to avoid all possible side effects, since Toral et al. (2018) have shown that the use of HTs as source texts facilitates the MT process mainly because of the decreased lexical variety. On the other hand, Freitag et al. (2019) recommend using both, albeit separated, original as well as HT source texts precisely in order to be able to take into account and better understand all effects.

Apart from the impact of translation direction, the impact of divergences from a source text in HT

[1] For example, in the WMT shared tasks, even texts written in an "external" original language were used extensively, e.g. English HTs from Czech texts were used as source for English-to-German MT systems.

[2] http://www.statmt.org/wmt19/

used as MT data has been investigated, too. The potential influence of different translation strategies and resulting divergences (shifts) on MT evaluation was discussed in (Popović, 2019), whereas Vyas et al. (2018) explored automatic identification of such divergences and their effects on MT training.

Texts translated by different translators Despite of a large body of work dealing with analysis of different translated texts in different contexts, there is, however, not much work about texts translated by different translators. Rubino et al. (2016) explored effects of translator's expertise to some extent, and reported that texts translated by students could be automatically distinguished from originals with higher accuracy than texts translated by professional translators. This indicates that the features of professional HTs are more similar to the features of originals. To the best of our knowledge, our work is the first attempt to systematically compare texts translated by different translators.

3 Data sets

For our experiments, we used three available parallel data sets involving three different language pairs and five translation directions: English→Croatian ($EnHr$), German↔French ($DeFr$) and English↔Finnish ($EnFi$). All data sets belong to the news domain and originate from the publicly available WMT shared tasks.

Ideally, each data set should have been designed specifically for one particular RQ, and created under the same conditions: each of the translators should have translated the same, sufficiently large source text. In addition, all source language texts should be originally written in that language, not being translated from some other language. Table 1 summarises the properties of our three data sets, and the following limitations can be noted:

* None of the data sets were specifically designed for one RQ: only the $EnHr$ data set is (almost) ideal for translation expertise (RQ1). The $DeFr$ is appropriate for both expertise (RQ1) and native language (RQ2), whereas the $EnFi$ data set is suitable for native language (RQ2).

* The $EnHr$ data set is, as mentioned above, almost ideal for exploring translation expertise, although one HT was generated from a different source text than the other three. The main drawback is that both source language texts were not written originally in English, but are HTs: they were translated from Czech in the framework of the WMT 2012 and WMT 2013 shared task However, this fact has no influence on the results of our experiment, because all HTs are coming from the same original language.

* The main limitation of the $DeFr$ data set is its small size: while in the other two data sets at least 1000 source sentences were translated by each translator, this data set ranges from 100 to 750 sentences. Another drawback is the lack of a common source text for all translators – therefore, different HTs represent a comparable corpus instead of a parallel corpus. The same limitation represents the main drawback of the $EnFi$ data set.

In addition, several domains/genres should ideally be covered, whereas all our data sets come from the news domain. Nevertheless, an ideal data set is, to the best of our knowledge, currently not available for any of our research questions. Therefore, we carried out our first experiments on the described available texts which, despite of their flaws, represent a good starting point for this research direction. The data sets are made publicly available for further research.[3] No personal information (such as name, gender, age, working place) about the translators are shared. The details and statistics of each of the texts used are presented together with the results in the corresponding sections.

4 Text features

The set of text features used in our experiments is inspired by the features frequently used in the literature (Baroni and Bernardini, 2006; Koppel and Ordan, 2011; Volansky et al., 2015; Toral, 2019). Although they are also motivated by two theoretical categories, simplification (Baker et al., 1993) and interference (Toury, 1979), they do not represent any of these categories exclusively. The choice of features is based on a hypothesis that the selected features might vary depending on the factors addressed in our work, namely translator's experience, native language and translation strategies.

For all features, punctuation marks were separated and counted as words. POS tags for all languages are generated by TreeTagger.[4] The features

[3] https://github.com/m-popovic/different-HTs

[4] https://www.cis.uni-muenchen.de/~schmid/

property	$EnHr$	$DeFr$	$EnFi$
RQ1: expertise	+	+	−
RQ2: translation direction and native language	−	+	+
same source language text for each translator	±	−	−
≥1000 sentences per translator	+	−	+
source language is the original language	−	+	+

Table 1: Properties of the three data sets; "−" denotes lack of a specific property.

are defined and calculated in the following way:

Sentence length: Number of words in each sentence of the text.

Some translators might tend to generate longer sentences in the target text than others. Some translators might keep the number of words in the translated sentences closer to the number of source text words than others.

Mean word length: The total number of characters in the text divided by total number of words.

Some translators might prefer longer (potentially more complex) words than others.

Lexical variety: The total number of distinct words in the text divided by the total number of words in the text.

$$lexVar = \frac{N(distinct\ words)}{N(words)} \quad (1)$$

Previous work has shown that vocabulary of HTs is generally less rich than vocabulary of originals. However, some translators might use more distinct words (a richer vocabulary) than others.

Morpho-syntactic variety: The total number of distinct POS tags in the text divided by the total number of words.

$$morphsynVar = \frac{N(distinct\ POS)}{N(words)} \quad (2)$$

Some translators might use more complex and/or more diverse grammatical structures than others. Some might keep the grammatical structure of translated sentences closer to the one of the source text than others.

Lexical density: The ratio between the total number of content words (adverbs, adjectives, nouns and verbs) and the total number of words.

$$lexDens = \frac{N(content\ words)}{N(words)} \quad (3)$$

tools/TreeTagger/

HTs have been found to have a lower percentage of content words than originals. However, some translators might use more content words than others.

5 Experimental set-up

For each feature, we calculate relative difference between the feature value of the original source text $f(source)$ and the feature value of its translation $f(ht)$.

$$\Delta(f) = \frac{f(source) - f(ht)}{f(source)} \quad (4)$$

The main benefits of reporting relative differences are:

- relative difference reduces impact of distinct source languages (language pairs);
- relative difference minimalises effects of using comparable instead of parallel HTs.

Table 2 shows an example of lexical varieties of two comparable HTs. The values of the two target language lexical varieties $f(ht)$ imply that the second HT is lexically richer. However, the reason for that difference might simply be the initially higher lexical variety of the second source text. Relative difference, though, clearly demonstrates that the second translation is lexically less rich and also closer to the source text.

	$f(ht)$	$f(source)$	$\Delta(f)$
source 1	0.721	0.434	**66.1%**
source 2	**0.832**	0.548 (!)	51.8%

Table 2: Example of analysing lexical varieties of two comparable HTs and advantage of using relative differences.

For each text and each feature, relative difference is calculated as average value over chunks of 100 sentences[5] (approximately 2000 words), similarly to some previous work (Volansky et al.,

[5]50 sentences (1000 words) for the $DeFr$ corpus due to the small size

2015). The purpose of averaging over small chunks is manifold: to make sure that the length of a text does not interfere with the feature values, to avoid issues related to the small size of some texts, and to further minimise the potential effects of using comparable instead of parallel translations.

For each of the research questions, the obtained values are reported and discussed in the following section. It is worth noting that the numbers differ between the data sets due to distinct properties of the language pairs. For example, relative difference between Finnish and English lexical and POS varieties are much larger than those between German and French.

We did not perform any text classification in this experiment, because the sizes of the currently available texts are not sufficient for training a classifier.

6 Results

6.1 RQ1: Influence of expertise and different cohorts

The $EnHr$ data set and the appropriate part of the $DeFr$ data set were used to examine the potential influence of different translator cohorts on text features. The statistics showing number of sentences and translator cohorts for both data sets is shown in Table 3. All translators were native speakers of the target language.

The $EnHr$ data set was created in the framework of the Abu-MaTran project.[6] A subset of the English test set[7] from WMT 2012 (1011 sentences) was translated into Croatian in two ways: professional translation and crowdsourcing via the CrowdFlower platform.[8] The options on the platform were configured in a way that enables the best possible translation quality: geography was limited to Croatia, and only the contributors on the top performance level were considered. In this way, 30 different crowdsourcing contributors participated in translation. In total, three HTs were created from this English source text: one by a professional translator and two by different crowd contributors. In a later phase of the project, 1000 English sentences[9] from the WMT 2013 were translated by a student, thus representing a third translator cohort, although as a comparable text.

The $DeFr$ data set was created for the WMT 2019 shared task. A subset of 1327 sentences was originally written in German and translated by translators with three different levels of expertise: student (326 sentences), professional translator (756 sentences), and specialist[10] (245 sentences).

6.1.1 Results on the *EnHr* data set

The main tendencies which can be observed in Table 4 are variations in sentence length and lexical variety, and to a lesser extent, in morphosyntactic variety. In addition, the features of the two crowd HTs are very similar, and more distinct than the features of the other two HTs. The sentence length indicates that the crowd produced shorter Croatian translations than the professional translator and the student. Higher lexical and morphosyntactic varieties are probably a consequence of a large number of different contributors which lead to a decrease in consistency. Here, it should be noted that a large lexical and/or grammatical variety as well as a large divergence from the source text are not necessarily positive.

Effects on automatic MT evaluation Since the $EnHr$ data set is the only one containing parallel (instead of comparable) HTs, it represents a perfect data set for testing the behaviour of automatic MT evaluation scores calculated on distinct reference translations. For this purpose, we translated the English source text by two online MT systems,[11] Google Translate[12] and Bing Translator.[13] We then calculated the widely used BLEU score (Post, 2018) and two recently proposed character-based metrics, F-score (Popović, 2015) and edit distance (Wang et al., 2016). All scores are calculated by comparing MT output with each of the HTs.

The resulting scores in Table 5 lead to different conclusions depending on the used reference HT. According to the professional HT, the Google MT output is substantially better than the Bing output in terms of all three evaluation metrics. If the first crowd HT is used as a reference, the differences between the two systems become small according to BLEU and chrF, whereas characTER even says that the Bing MT output is better. A similar tendency can be observed if the student HT is

[6] https://www.abumatran.eu/
[7] HT from Czech, as mentioned in Section 3
[8] http://crowdflower.com/
[9] also HT from Czech
[10] not a professional translator by vocation, but experienced
[11] in November 2019
[12] https://translate.google.com/
[13] https://www.bing.com/translator

data set	parallel text	translator	translation expertise	number of sentence pairs
$EnHr$	2012 en→hr	$\mathrm{T}hr_1$	professional	1011
	2012 en→hr	$\mathrm{T}hr_2$	crowd	
	2012 en→hr	$\mathrm{T}hr_3$	crowd	
	2013 en→hr	$\mathrm{T}hr_4$	student	1000
$DeFr$	2019 de→fr	$\mathrm{T}fr_1$	student	326
	2019 de→fr	$\mathrm{T}fr_2$	specialist	245
	2019 de→fr	$\mathrm{T}fr_3$	professional	756

Table 3: Characteristics of the texts used to examine the influence of translation expertise: language pair, translator, translator's expertise and number of sentences.

EnHr **en→hr**	translator expertise	$\mathrm{T}hr_1$ prof.	$\mathrm{T}hr_2$ crowd	$\mathrm{T}hr_3$ crowd	$\mathrm{T}hr_4$ stud.
$\frac{SRC(en)-HT(hr)}{SRC(en)}$	Δ(sentence length)	**8.06**	**12.8**	**13.4**	**11.3**
	Δ(word length)	-13.7	-14.3	-15.0	-14.8
	Δ(lexical variety)	**-32.6**	**-40.3**	**-41.6**	**-35.6**
	Δ(POS variety)	**-413**	**-426**	**-423**	**-406**
	Δ(lexical density)	51.9	51.6	51.8	53.1

Table 4: Relative differences (%) between features of the original texts and features of the translated texts for English→Croatian texts translated by translators with different expertises: professional, crowd and student.

used, albeit the comparison is not completely appropriate since the source text is different. If the second crowd HT is used, the BLEU score of the Bing output becomes slightly better, the characTER score becomes substantially better, whereas the chrF score is slightly worse than Google.

The fact that automatic scores calculated on different reference translations are different is, of course, nothing new. However, here we point out that translator cohort providing the reference HT can have influence on the scores and perceptions of systems' quality, and therefore represents a factor which should be taken into account in MT evaluation.

6.1.2 Results on the *DeFr* data set

Table 6 shows the text features of the $DeFr$ data set. In spite of differences between this corpus and the $EnHr$ corpus in terms of expertise levels, languages, as well as comparable HTs instead of parallel HTs, the same general tendencies can be observed, namely variations in sentence length, lexical variety and morpho-syntactic variety. The sentences in the professional HT are longest and the lexical variety is highest, which could be intuitively expected – professional translators tend to divert more from the source language and to use richer vocabulary. Morpho-syntactic variety, however, is highest in the specialist HT, although not much higher than in the other two. All the findings indicate that translation expertise has influence on sentence length, lexical and morpho-syntactic variety, however a deeper analysis is needed in the future to identify the nature of these differences.

Lexical density, however, varies only in the $DeFr$ data set, especially for the specialist's translation. This feature should certainly be analysed further in order to determine whether the variations are related to the translator expertise, or maybe to some other factors such as distinct nature of the language pair, translator's individual preferences, etc.

6.2 RQ2: Influence of native language and translation direction

The differences between native and non-native HTs were analysed on appropriate portions of the $DeFr$ and $EnFi$ data sets. The statistics of the texts used are shown in Table 7. As already mentioned in Section 3, both data sets contain comparable HTs.

The $DeFr$ texts, created for the WMT 2019 shared task, enable two ways of investigating influence of (non-)native language. One is to compare two translation directions of one translator: a French native specialist $\mathrm{T}fr_2$ was translating in both directions: from French into German (from their native language) and from German into French (into their native language). Another way is to compare two translators working on the same translation direction: a French native specialist $\mathrm{T}fr_2$ and a German native specialist $\mathrm{T}de_1$ both were translating from French into German.

The $EnFi$ data set enables only the first type

***EnHr*, en→hr**	BLEU ↑		chrF ↑		characTER ↓	
reference	Google	Bing	Google	Bing	Google	Bing
Thr_1 (professional)	41.9	34.9	65.5	60.3	30.3	33.4
Thr_2 (crowd)	32.9	32.6	59.8	59.0	**34.1**	**33.4**
Thr_3 (crowd)	**29.5**	**29.6**	57.6	57.4	**36.2**	**35.0**
Thr_4 (student)	34.7	31.2	58.8	57.5	**35.9**	**35.7**

Table 5: Three automatic evaluation scores (BLEU, chrF and characTER) for English-to-Croatian on-line MT systems calculated on reference translations produced by translators with different expertises: professional, crowd and student.

DeFr **de→fr**	translator expertise	Tfr_1 stud.	Tfr_2 spec.	Tfr_3 prof.
$\frac{SRC(de)-HT(fr)}{SRC(de)}$	Δ(sentence length)	**-21.3**	**-23.4**	**-26.1**
	Δ(word length)	10.6	11.4	10.9
	Δ(lexical variety)	**12.8**	**10.3**	**14.8**
	Δ(POS variety)	**39.8**	**40.9**	**38.7**
	Δ(lexical density)	**-10.2**	**-5.62**	**-13.2**

Table 6: Relative differences (%) between features of the original texts and features of the translated texts for German→French texts translated by translators with different expertises: student, specialist and professional.

of analysis, namely comparison of two translation directions done by one translator. It contains three HTs produced by a Finnish native professional: one English into Finnish (into their native language) translation and two Finnish to English (from their native language) translations.

Table 8 presents text features for all native and non-native HTs. **Texts translated by one translator in two different translation directions** are compared in Table 8(a). The following general tendencies can be observed for both translators and both language pairs: sentence length, word length and lexical variety substantially differ depending on the translation direction. Word length and lexical variety are higher when translating into the native language, indicating that the translators tend to choose longer words more often and to use a richer vocabulary in their native language, as intuitively can be expected. As for sentence length, the differences tend in opposite directions: for $DeFr$, the length of non-native HTs is closer to the source text (which can be intuitively expected), whereas for $EnFi$ is the other way round. The reason might lay in sheer differences between the two language pairs, which should be investigated in future work. Deeper analysis of reasons and underlying phenomena is also needed for POS variety and lexical density, because the tendencies are very different for the two language pairs: in $DeFr$ texts, lexical density varies whereas there are no large differences in POS variety, and in $EnFi$ texts is the other way round.

Table 8(b) shows the features of **texts translated from French into German by two translators with different native languages**. It can be seen that the variations in sentence length, word length and lexical variety observed in Table 8(a) are confirmed. Furthermore, word length and lexical variety are again higher in the native translations. Sentence length of the non-native HT is closer to the source language, same as the other $DeFr$ non-native HT presented in Table 8(a) – this also indicates that the reason for the opposite tendency observed on the $EnFi$ language pair might indeed be the different nature of the language pair itself. In any case, a detailed analysis is definitely necessary, as well as for morpho-syntactic variety and lexical density.

Despite the fact that certain tendencies should be investigated further, it can be noted that native and non-native translated texts generally exhibit different traits, especially regarding sentence length, word length and lexical variety. Therefore, the native language of the translator should also be taken into account for MT evaluation.

7 Conclusions

This work presents results of a set of computational analyses on three data sets containing three language pairs and five translation directions with the aim of finding out whether different human translations exhibit different traits. Despite certain limitations, our findings represent a good base for analysing different human translations.

The main contribution of this work is empirical, showing that each of the investigated factors has certain influence on the features of translated texts. Sentence length and lexical variety are affected by

data set	parallel text	translator	translation direction	number of sentence pairs
$DeFr$	2019 de→fr	Tfr_2	into native	245
	2019 fr→de	Tfr_2	from native	235
	2019 fr→de	Tde_1	into native	100
$EnFi$	2017 en→fi	Tfi_1	into native	1502
	2017 fi→en	Tfi_1	from native	1500
	2019 fi→en	Tfi_1	from native	1996

Table 7: Characteristics of the texts used to examine the influence of native language and translation direction: language pair, translator, translation direction, and number of sentences.

(a) one translator, two translation directions

DeFr **Tfr$_2$**	translation direction	de→fr (into native)	fr→de (from native)
$\frac{SRC-HT}{SRC}$	Δ(sentence length)	**-23.4**	**0.21**
	Δ(word length)	**11.4**	**-4.96**
	Δ(lexical variety)	**10.3**	**-2.82**
	Δ(POS variety)	40.9	-41.3
	Δ(lexical density)	**-5.62**	**19.0**

EnFi **Tfi$_1$**	translation direction	en→fi (into native)	fi→en (from native)	(from native)
$\frac{SRC-HT}{SRC}$	Δ(sentence length)	**21.5**	**-51.4**	**-46.8**
	Δ(word length)	**-55.5**	**36.2**	**35.0**
	Δ(lexical variety)	**-46.7**	**39.4**	**36.3**
	Δ(POS variety)	**-393**	**79.8**	**79.3**
	Δ(lexical density)	-26.5	24.3	26.2

(b) one translation direction, two translators

DeFr **fr→de**	translator	Tde_1 (into native)	Tfr_2 (from native)
$\frac{SRC(fr)-HT(de)}{SRC(fr)}$	Δ(sentence length)	**5.46**	**0.21**
	Δ(word length)	**-9.64**	**-4.96**
	Δ(lexical variety)	**-3.42**	**-2.82**
	Δ(POS variety)	**-24.4**	**-41.3**
	Δ(lexical density)	**22.9**	**19.0**

Table 8: Relative differences (%) between features of native and non-native HTs; one translator working on two translation directions (a) and two translators working on one translation direction (b).

all factors, whereas word length varies depending on native language. As for POS variety and lexical density, a deeper analysis is needed to understand the observed tendencies. While we believe that the trends observed in the reported results are not incidental, more research is needed to find linguistic explanations. Our study is based on rather superficial text features at word and POS level – therefore, for future work, different HTs should be analysed in depth, including over- or under-using particular words, collocations and POS categories, as well as presence or absence of different types of translation shifts and semantic divergences. Furthermore, as described in Section 3, this study is carried out on sub-optimal data sets – providing and investigating larger data sets containing parallel HTs generated from the same source text is necessary. More data will also enable another line of work, namely automatic discrimination between different HTs.

More (ideal) data will also enable better analysis of potential effects on human and automatic MT evaluation. Nevertheless, even the presented preliminary results suggest that it is important to specify which kind of HTs were used for MT evaluation, especially for evaluations which involve comparing human and machine translation quality. As MT quality improves, such comparisons are becoming more and more frequent, and are also becoming a part of WMT shared tasks – at the WMT 2019 shared task ((Barrault et al., 2019), Section 3.8), for the German-English language pair it is reported that "many systems are tied with human performance", as well as that "Facebook-FAIR system achieves super-human translation performance". For this type of evaluation, we highly

recommend that researchers/evalutaors specify the details about the HTs used.

8 Acknowledgments

The ADAPT SFI Centre for Digital Media Technology is funded by Science Foundation Ireland through the SFI Research Centres Programme and is co-funded under the European Regional Development Fund (ERDF) through Grant 13/RC/2106.

Special thanks to Maarit Koponen, Antonio Toral, Barry Haddow, Loïc Barrault, Franck Burlot, Tereza Vojtěchová and Mārcis Pinnis for all invaluable information and support.

References

Ahrenberg, Lars. 2017. Comparing machine translation and human translation: A case study. In *Proceedings of the 1st Workshop on Human-Informed Translation and Interpreting Technology (HiT-IT 2017)*, pages 21–28, Varna, Bulgaria, September.

Baker, Mona, Gill Francis, and Elena Tognini-Bonelli. 1993. Corpus linguistics and translation studies: Implications and applications. *Text and Technology: in Honour of John Sinclair*, pages 233–250.

Baroni, Marco and Silvia Bernardini. 2006. A new approach to the study of translationese: Machine-learning the difference between original and translated text. *Literary and Linguistic Computing*, 21(3):259–274, September.

Barrault, Loïc, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 Conference on Machine Translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (WMT 2019)*, pages 1–61, Florence, Italy, August.

Castilho, Sheila, Natalia Resende, and Ruslan Mitkov. 2019. What influences the features of post-editese? a preliminary study. In *Proceedings of the 2nd Workshop on Human-Informed Translation and Interpreting Technology (HiT-IT 2019)*, pages 20–28, Varna, Bulgaria, September.

Chesterman, Andrew. 2004. Beyond the particular. In *Translation universals: Do they exist?*, pages 33–50. John Benjamins.

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 2016)*, pages 106–114, Riga, Latvia.

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 Studie*, 16:89–103.

Farrell, Michael. 2018. Machine translation markers in post-edited machine translation output. In *Proceedings of the 40th Conference on Translating and the Computer (TC40)*, pages 50–59, London, UK, November.

Freitag, Markus, Isaac Caswell, and Scott Roy. 2019. APE at Scale and Its Implications on MT Evaluation Biases. In *Proceedings of the Fourth Conference on Machine Translation (WMT 2019)*, pages 34–44, Florence, Italy, August.

Koppel, Moshe and Noam Ordan. 2011. Translationese and its dialects. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT 2011)*, pages 1318–1326, Portland, Oregon, June.

Kurokawa, David, Cyril Goutte, and Pierre Isabelle. 2009. Automatic detection of translated text and its impact on machine translation. In *In Proceedings of MT Summit XII*, pages 81–88, Ottawa, Canada, August.

Läubli, Samuel, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? a case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018)*, pages 4791–4796, Brussels, Belgium, October-November.

Lembersky, Gennadi, Noam Ordan, and Shuly Wintner. 2013. Improving statistical machine translation by adapting translation models to translationese. *Computational Linguistics*, 39(4):999–1023.

Popović, Maja. 2015. chrF: character n-gram f-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal, September.

Popović, Maja. 2019. On reducing translation shifts in translations intended for MT evaluation. In *Proceedings of Machine Translation Summit XVII*, pages 80–87, Dublin, Ireland, 19–23 August.

Post, Matt. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium, October.

Rabinovich, Ella and Shuly Wintner. 2015. Unsupervised identification of translationese. *Transactions of the Association for Computational Linguistics*, 3:419–432, December.

Rabinovich, Ella, Sergiu Nisioi, Noam Ordan, and Shuly Wintner. 2016. On the similarities between

native, non-native and translated texts. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, pages 1870–1881, Berlin, Germany, August.

Rubino, Raphael, Ekaterina Lapshinova-Koltunski, and Josef van Genabith. 2016. Information density and quality estimation features as translationese indicators for human translation classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2016)*, pages 960–970, San Diego, California, June.

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 3rd Conference on Machine Translation (WMT 2018)*, pages 113–123, Belgium, Brussels, October.

Toral, Antonio. 2019. Post-editese: an exacerbated translationese. In *Proceedings of Machine Translation Summit XVII*, pages 273–281, Dublin, Ireland, 19–23 August.

Toury, Gideon. 1979. Interlanguage and its manifestations in translation. *Meta*, 24(2):223–231.

Vanmassenhove, Eva, Dimitar Shterionov, and Andy Way. 2019. Lost in translation: Loss and decay of linguistic richness in machine translation. In *Proceedings of Machine Translation Summit XVII*, pages 222–232, Dublin, Ireland, 19–23 August.

Volansky, Vered, Noam Ordan, and Shuly Wintner. 2015. On the features of translationese. *Digital Scholarship in the Humanities DSH*, 30(1):98–118, April.

Vyas, Yogarshi, Xing Niu, and Marine Carpuat. 2018. Identifying semantic divergences in parallel text without annotations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2018)*, pages 1503–1515, New Orleans, Louisiana, June.

Wang, Weiyue, Jan-Thorsten Peter, Hendrik Rosendahl, and Hermann Ney. 2016. Character: Translation edit rate on character level. In *Proceedings of the First Conference on Machine Translation*, pages 505–510, Berlin, Germany, August.

Zhang, Mike and Antonio Toral. 2019. The effect of translationese in machine translation test sets. In *Proceedings of the 4th Conference on Machine Translation (WMT 2019)*, pages 73–81, Florence, Italy, August. Association for Computational Linguistics.