Uncovering Machine Translationese Using Corpus Analysis Techniques to Distinguish between Original and Machine-Translated French

Orphée de Clercq, Gert de Sutter, Rudy Loock, Bert Cappelle, Koen Plevoets

Uncovering Machine Translationese Using Corpus Analysis Techniques to Distinguish between Original and Machine-Translated French

Orphée De Clercq[1] *Gert De Sutter*[2] *Rudy Loock*[3]

Bert Cappelle[4] *Koen Plevoets*[5]

Address: [1,2,5] Department of Translation, Interpreting and Communication, Ghent University, Groot-Brittanniëlaan 45, 9000 Ghent, Belgium;
[3,4] Université de Lille, CNRS "Savoirs, Textes, Langage" CNRS Research Unit, 3 Rue du Barreau, 59650 Villeneuve-d'Ascq, France
E-mail: [1]orphee.declercq@ugent.be; [2]gert.desutter@ugent.be; [3]rudy.loock@univ-lille.fr; [4]bert.cappelle@univ-lille.fr; [5]koen.plevoets@ugent.be
Correspondence: Orphée De Clercq
Citation: De Clercq, Orphée, De Sutter, Gert, Loock, Rudy, Cappelle, Bert and Koen Plevoets. 2021. "Uncovering machine translationese using corpus analysis techniques to distinguish between original and machine-translated French." *Translation Quarterly* 101: 21-45.

Abstract

This paper investigates the linguistic characteristics of English to French machine-translated texts in comparison with French original, untranslated texts in order to uncover what has been called "machine translationese". In the same vein as corpus-based translation studies which have focused on human-translated texts, and using a corpus-based statistical approach (Principal Component Analysis), we analyzed a ca. 1.8-million-word corpus of English to French translations of press texts, corresponding to the output of four machine translation systems: one statistical (SMT) and three neural (NMT) systems, namely DeepL, Google Translate, and the European Commission's eTranslation MT tool, in both its SMT and NMT versions. In particular, to complement a previous study on language-specific features in French (e.g. derived adverbs, existential constructions, coordinator et, preposition avec), a series of language-independent linguistic features were extracted for each text in our corpus, ranging from superficial text characteristics such as average word and sentence length to frequencies of closed-class lexical categories and measures of lexical diversity. Our results, which compa-

re the machine-translated data with a corpus of French untranslated data, allow us to uncover linguistic features in French machine-translated texts that clearly deviate from the observed norms in original French (e.g. average sentence length, n-gram features, lexical diversity), and which might serve as information for the post-editing process in order to optimize translation quality.

1. Introduction

Since the advent of neural machine translation (NMT) systems (Forcada 2017), it has become clear that the technology is disruptive and brings a lot of changes to the translation industry, both in terms of translation process and business model. In particular, as opposed to statistical machine translation (SMT), current NMT systems have started to provide output whose fluency can be quite impressive, although sometimes at the expense of accuracy or fidelity to the source text (e.g. Bojar et al. 2016; Macken et al. 2019). Such fluency has even led some researchers to claim that MT systems have reached "human parity" (Hassan et al. 2018), although such claims have been reassessed since (Toral et al. 2018). Nevertheless, the improved quality has led to the promotion of NMT in the field of translation (Daems and Macken 2019).

This makes it all the more crucial for translators to define their added value over the machine: they should develop their MT literacy, a concept defined by Bowker and Ciro (2019) for non-professionals, that is, they need to know what the machine can(not) do, what the difference is between human translations and MT output, and what to focus on during the post-editing (PE) process.

With NMT, the evaluation of MT systems has become a central issue, not only for the industry but also for translation training. The focus on fluency makes errors more difficult to identify (e.g. Castilho et al. 2017a, 2017b; Yamada 2019), and translators need to be provided with useful information for the PE process. A lot of debates have been taking place on the best way to assess MT output quality: use of metrics, human evaluation, or a linguistic evaluation with a corpus-based approach. This paper focuses on such a linguistic evaluation of MT output, through the analysis of English to French machine-translated texts produced by four different MT systems, in comparison with original, untranslated French data. Our analysis aims to explore different aspects of what has been called "machine translationese" (see e.g. Daems et al. 2017) by comparing machine-translated with original texts, relying on a corpus-based approach typical of corpus-based translation studies, with analyses carried out on a series of texts rather than a series of isolated sentences.

To this purpose a corpus of press texts was collected comprising both original (French) French and (British) English text material. The English data was translated into French using four different MT systems: DeepL, Google Translate – both NMT – and the European Com-

mission's eTranslation tool, in both the SMT and NMT flavor. This allowed us to compare the frequencies of a series of linguistic features in original vs. machine-translated French, with the same methodology and statistical techniques that have proven capable of distinguishing between student and professional translations (De Sutter et al. 2017). The final aim is to define the 'gap' that exists between machine-translated texts and the norms expected in untranslated texts, providing information on how to improve translators' invisibility as expected by today's market, thanks to a list of elements to focus on during the PE process.

Specifically, this paper aims to complement a previous study (Loock 2018, 2020), which analyzed the same data (with the exception of Google Translate output) by focusing on language-*specific* linguistic features (see section 2.2 for a summary and list of the features). Here our focus is on language- *independent* features like average word or sentence length, frequencies of part-of-speech (POS) tags, or frequencies of n-grams. These features are exploratively analyzed with Principal Component Analysis and the differences between original and machine-translated French are then tested by means of ANOVA. The analysis is therefore more sophisticated than in Loock (2018, 2020) and also includes the most famous publicly available MT system.

The remainder of this paper is structured as follows. In section 2 we describe related work on machine translationese within research on the quality of MT output and provide a summary of Loock (2018, 2020), of which the current study is an extension. Section 3 describes our methodology: corpus material, feature extraction, and statistical technique. Section 4 is dedicated to the presentation and discussion of the results, first for a general comparison between French machine-translated and untranslated texts, then for a finer-grained comparison of relevant linguistic features, and finally for a possible link with interference from the English source texts.

2. Related work

2.1 MT output evaluation and machine translationese

A lot of research has been devoted to the evaluation of MT output (see Moorkens et al. 2018 for a recent overview), in particular since the advent of NMT. Alongside metrics like BLEU (BiLingual Evaluation Understudy; Papineni et al. 2002) for example, researchers have relied on human evaluations to try and tackle the limits of automatic evaluation (see e.g. Babych 2014)[1]. For example, MT output has been evaluated by identifying and classifying errors (e.g. Federico et al. 2013; Van Brussel et al. 2018), measuring the amount of post-editing effort (e.g. Bentivogli et al. 2016), or ranking machine-translated texts by (non-)professionals (e.g. Bojar et al. 2015). Researchers have also focused on the identification of linguistic differences between machine-translated texts and original, untranslated texts in the same language.

Following the path mapped out by corpus-based translation studies (CBTS) with Baker (1993) as a starting point, which allowed for the identification of linguistic differences between original and (human) translated texts (see Laviosa 2002; Olohan 2004 or De Sutter et al. 2017 for a series of quantitative studies), some studies have identified linguistic differences between untranslated language and machine-translated language, post-edited or not, for series of sentences (e.g. Isabelle et al. 2017) or full texts (e.g. Vanmassenhove et al. 2019). In the latter case, machine-translated texts are gathered as electronic corpora, to be investigated with the quantitative methods of corpus linguistics. Just as some of the analyses of human-translated language in CBTS led to the identification of translationese (Gellerstam 1986) [2], the observation of machine-translated texts has led to the identification of so-called "machine translationese" for raw MT output and "post-editese" for post-edited MT output (MTPE). For example, Vanmassenhove et al. (2019) have shown that MT texts show lesser lexical variety than both original and human-translated texts for English to French and English to Spanish translations; Lapshinova-Koltunski (2015) has investigated English to German translations and has, in the same vein as what has been done for the analysis of human-translated texts, investigated the possible influence of so-called translation universals like simplification, explicitation, and normalization, by measuring lexical density/variety, the frequency of cohesion markers or specific grammatical categories (nouns vs. verbs). Similarly, Daems et al. (2017) and Toral (2019) have analyzed MTPE texts and shown the existence of post-editese, qualified by Daems et al. (2017) as "exacerbated translationese". In the present study, our focus is on raw, non post-edited MT output, and our aim is to check for the existence of machine translationese.

2.2 Language-specific vs. language-independent linguistic features

In order to uncover machine translationese or post-editese, researchers can focus on language-specific or language-independent linguistic features. For example, lexical variety or density in Lapshinova-Koltunski (2015) and average sentence/word length in Daems et al. (2017) are language-independent features, while Isabelle et al. (2017) focus on language-specific features for the evaluation of French to English MT output (e.g. verb-tense concordance, insertion of words like *fact or how*).

Our study investigates language-independent features (see complete list in section 3.2) in EN-FR machine-translated texts, as it is meant to complement a previous study on English to French machine-translated texts (Loock 2018, 2020) which used the same data (with the exception of the Google Translate subcorpus) and investigated specific linguistic features in French: the use of the hypernyms *chose and dire* ('thing' and 'say'), the coordinator *et* ('and'), the preposition *avec* ('with'), derived adverbs ending in *-ment* (the equivalent of *-ly* adverbs), and *il y* a existential constructions (the equivalent of *there is/are* constructions). The analysis of the EN-FR machine-translated texts (obtained by means of DeepL and eTranslation in both its SMT and NMT versions) has shown that, on an almost systematic basis, these specific lin-

guistic features show highly significant differences between original, untranslated French and machine-translated French from English, with much higher frequencies in machine-translated texts. These French linguistic features were selected as they are considered to be translational equivalents of the corresponding English items but these items' use in original English and original French shows differences in terms of frequencies. As the items are more frequent in original English than original French, one would expect the differences observed in machine-translated French texts to be the result of direct transfers between the English source texts and the French translated texts. However, the qualitative analysis carried out in Loock (2020) shows that this is not the case. Source language interference cannot fully explain the data, as we also notice differences in frequencies between the English source texts and the French machine-translated texts: for example, the frequency of *il y* a existential constructions in machine-translated French, higher than in untranslated French, is lower than that of *there is/are* constructions in the English source texts, suggesting that only some of them are translated literally (this is confirmed by the qualitative analysis of a sample of the corpus in Loock 2020).

3. Method

Using a corpus-based statistical approach, our objective is to investigate the existence of "machine translationese" for the language pair English-French. This approach consists of three steps. First, original French and English press texts are collected, after which the English texts are machine-translated using four well-known MT engines (section 3.1). Next, all corpora are preprocessed (including tokenization, lemmatization and part-of-speech tagging) and subsequently a series of linguistic language-independent features are extracted (section 3.2). In the third step, multivariate statistical analysis techniques are performed and the output is analyzed (section 3.3).

3.1 Data collection

The data used for this study contains three kinds of texts: (i) original texts written in (British) English, (ii) their translations into French by means of 4 different MT systems, and (iii) untranslated texts written in (French) French. Both series of original texts (i/iii) are extracted from the TSM press corpus (*Traduction Spécialisée Multilingue* corpus), an open-ended corpus compiled at the University of Lille for a comparative grammar class in a master's programme (Loock 2019). The corpus is a comparable corpus containing original, untranslated press texts taken from quality newspapers in British English (e.g. *The Guardian, The Observer, The Times*), American English (e.g. *The New York Times, The Wall Street Journal*), and (French) French (e.g. *Le Monde, Libération, La Voix du Nord*), with different news domains being covered: business and finance, crime, culture, environment, health, international news,

politics, science & technologies, sports and travel. At the time the current study was initiated, the TSM corpus contained ca. 1.6 million words (2.4 million words today). Table 1 provides a description of the version of the TSM corpus used for the present study. All French texts were used (1,440 texts amounting to 833,590 words); for English the British English subcorpus was selected (490 texts amounting to 374,326 words).

Table 1: Content of the TSM press corpus

	Ori US_EN	Ori GB_EN	Ori FR
Business & Finance	27 487	6 136	64 361
Crime	44 315	43 710	120 343
Culture	30 570	46 839	107 080
Environment	41 500	32 367	101 924
Health	34 790	28 170	78 022
International News	33 767	29 168	91 147
Politics	45 840	46 901	127 500
Science & Technologies	45 269	47 213	94 391
Sports	45 156	43 766	125 033
Travel	40 748	50 056	108 493
Total #tokens	**389 442**	**374 326**	**833 590**
#texts	**437**	**490**	**1 440**
GRAND TOTAL	**1 597 358**		

As far as machine-translated texts are concerned, the 490 British English texts were translated using four translation engines: DeepL, Google Translate – two commercial neural engines that are freely available online – and the engine developed by the European Commission's Directorate-General for Translation, called eTranslation, both in its SMT and NMT versions. DeepL [3] is trained on the corpus used for the Linguee website[4] and has become known for the quality of the target language, sometimes at the expense of accuracy (Bojar et al. 2016). Google Translate [5] is probably the most well-known generic MT tool and is frequently the object of scientific studies on the quality of MT output. Both tools have been providing internet users with neural machine translations since NMT went mainstream (around 2016). The eTranslation tool[6], both in its SMT and NMT flavors, has been designed for internal use at the European Commission and is not available to the general public,[7] although public administrations as well as small and medium-sized enterprises can currently make use of it, with September 2018 marking the arrival of the NMT version for the EN-FR language pair. In spite of its confidential nature, eTranslation has been the object of a few studies (Macken et al. 2020; Rossi and Chevrot 2019; Loock 2020).

The translations were retrieved in spring 2018 for DeepL and eTranslation SMT, Decem-

ber 2018 for eTranslation NMT, and July-August 2019 for Google Translate. Each of the 490 texts was translated individually, by copying/pasting the text online or by uploading the different files. Table 2 provides some corpus statistics of the machine-translated data, the British English source texts as well as a specification of the size of the comparable original French corpus used for the present study.

Table 2: Content of the corpus used for this study

Subcorpus		#texts	#tokens	Abbreviation
Original French		1 440	833 590	ORI_FRA
Machine-translated French	DeepL	490	442 439	DeepL
	eTranslation NMT	490	451 704	eNMT
	eTranslation SMT	490	445 914	eSMT
	Google Translate	490	431 297	GoogT
Original English (British)		490	374 326	SCR_ENG
GRAND TOTAL		3 890	2 979 270	

It should be noted that the original texts from the TSM press corpus belong to the press genre, while none of the 4 MT tools used have been trained or optimized for the translation of such texts (DeepL and Google Translate are generic MT tools; eTranslation has been trained on institutional data). This is of course a limitation of our study, since none of the 4 MT systems have been trained to translate press texts specifically, meaning the tools are not fully fit-for-purpose.

3.2 Data processing

A number of language-independent features have been extracted from the different subcorpora. For this extraction, it was crucial to linguistically preprocess all three corpora. This preprocessing consisted of three steps: tokenization, lemmatization and part-of-speech (POS) tagging. The LeTs preprocessing toolkit (Van de Kauter et al. 2013) was used for this purpose. The complete list of 22 features is presented in Table 3.

As can be derived from this table, the list contains two basic readability features (average word and sentence length), measures of lexical creativity and originality (e.g. type-token ratio, lexical density, hapax legomena), basic frequency information on different part of speech categories (lexico-grammatical features) and features indicating the degree of syntagmatic patterning or formulaicity (3- and 4-grams). All features are extracted at the text level; the legomena and ngram features are calculated by comparing an individual text with a background corpus. For example, for the legomena features we count how many French words in a certain text also occur one (hapax), two (dis) or three (tris) times in the entire French corpus used for this study. For the ngram features we check the number of combinations of three (3-gram) or

Table 3: All 22 language-independent features which were extracted from every text.

Feature type	Feature name	Abbreviation
Readability	Average sentence length	ASL
	Average word length	AWL
Lexical creativity	Lexical density	Den
	Type-token ratio	TTR
	Hapax legomena	Hapax
	Dis legomena	Dis
	Tris legomena	Tris
Lexico-grammatical	Frequency of common nouns	NOM
	Frequency of proper nouns	NAM
	Frequency of adjectives	ADJ
	Frequency of adverbs	ADV
	Frequency of verbs	VER
	Frequency of pronominals	PRO
	Frequency of determiners	DET
	Frequency of foreign words	FW
	Frequency of interjections	INT
	Frequency of numerals	NUM
	Frequency of prepositions & conjunctions	KON_PRP
Formulaicity	3-grams (word)	N3_wrd
	3-grams (POS)	N3_pos
	4-grams (word)	N4_wrd
	4-grams (POS)	N4_pos

four (4-grams) words or part-of-speech categories belonging to the 100 most frequent combinations in the corpus. Please note that when we refer to words for the legomena and ngram features, we actually mean the lemmas. Regarding the lexico-grammatical features it should be noted that the morphosyntactic categories prepositions and conjunctions were merged into one feature as we could not unequivocally distinguish these part-of-speech categories in the tagsets of both languages.

3.3 Statistical analyses

All language-independent features were extracted by means of custom-made Python scripts. This resulted in a data matrix in which every row contains the numerical information of the 22 features with respect to a given text. Every text is thus represented as a feature vector consisting of the scores of 22 linguistic features. Moreover, the origin of the text is also taken into

account – original French, machine-translated French using either DeepL, GoogleT, eNMT or eSMT or original English – leading to a 23-dimensional vector.

After having extracted this quantitative information from the corpora, we used Principal Component Analysis (PCA) to inspect the correlation structure of our data matrix in a lower-dimensional structure (see the seminal work by Baayen (2008) for more information, and Jenset and McGillivray (2012) or Evert and Neumann (2017) for examples of use of this methodology to uncover differences between original and translated language). For ease of presentation, we will present only two-dimensional plots in the remainder of this paper. These visualizations will reveal whether original and machine-translated French differ from each other, which could hint at machine translationese, and if so, whether all MT engines present the same picture. By also incorporating the corpus of British English source texts, possible shining-through from the source texts might also become apparent. This first explorative analysis is subsequently corroborated by an ANOVA of each linguistic feature, where the difference between original French, machine-translated French (with each of the four MT engines) and English is statistically tested. We will only report on the ANOVAs of the features which yield explicit differences.

4. Results and discussion

As mentioned in section 3, all texts from each subcorpus were first preprocessed, after which 22 language-independent features were extracted. Next, PCA was used to analyze the data. The results of this PCA are presented in Figure 1.

4.1 PCA results

Before interpreting this plot, one should note that the abbreviations printed in black represent the 22 linguistic features (see Table 3) and that each colored item represents a text coming from one of the different subcorpora. The numerical values on the x- and y-axis do not have a straightforward interpretation; what is meaningful, however, is the relative position of the different items vis-à-vis each other and vis-à-vis the linguistic features in the plot: the closer two items are, the more similar their linguistic characteristics (and vice versa); when a text is close to a given linguistic feature this means that the feature is clearly present in this text.

Given the number of texts, subcorpora, and features, this biplot is rather difficult to read. However, what immediately draws the attention is that we can distinguish between the English source texts (yellow) at the top, the original French texts (red) in the middle and the machine-translated French texts (all the other colors) more at the bottom. This means that there exist some clear differences between the different corpora, and that the linguistic characteristics of French machine-translated texts show some differences with both original French texts and

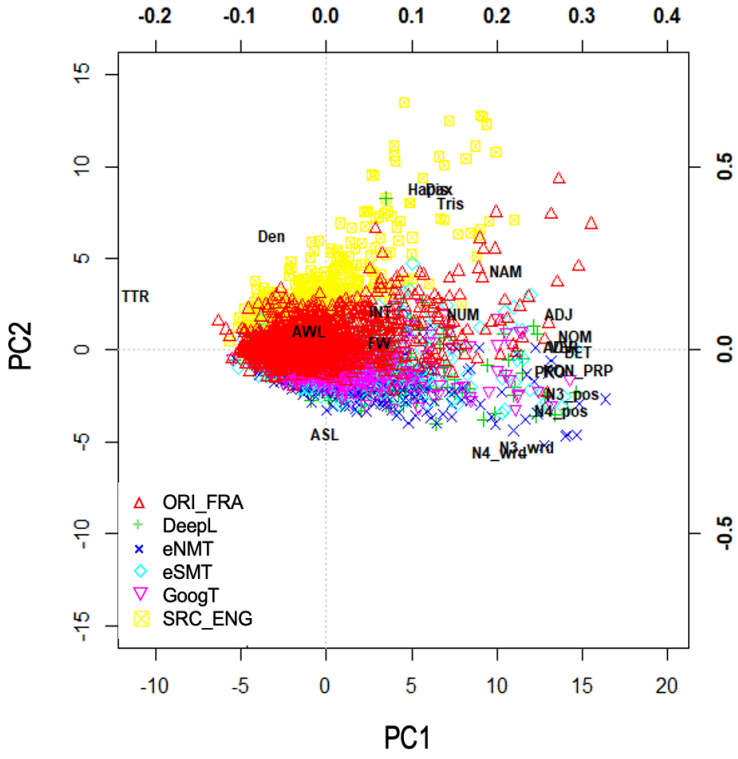

Figure 1: Biplot of the PCA for the original French (red triangle), the machine-translated French (green plus sign: DeepL, darkblue cross: eNMT, light blue diamond: eSMT, pink triangle: GoogT) and the British-English source texts (yellow check-marked square)

English source texts.

In order to better focus on the variation between the different varieties of French (untranslated texts and machine-translated texts for the different MT tools), we provide Figure 2, which depicts the same PCA analysis but with the visualization of the English source texts left out.

Looking at the original versus machine-translated French texts, there is quite some overlap, represented by the big colored blobs in the middle, though overall we also observe that certain features seem to pull down the machine-translated texts towards the right bottom, indicating that there does exist such a thing as machine translationese. If we look closer at which features cause this we see that it is mostly due to the average sentence length (ASL) and ngram features (N3_wrd, N4_wrd, N3_pos, N4_pos).

4.2 ANOVA analyses

This is where ANOVA analyses can shed more insights, as these test for each feature individually whether there is a significant difference between the different settings.

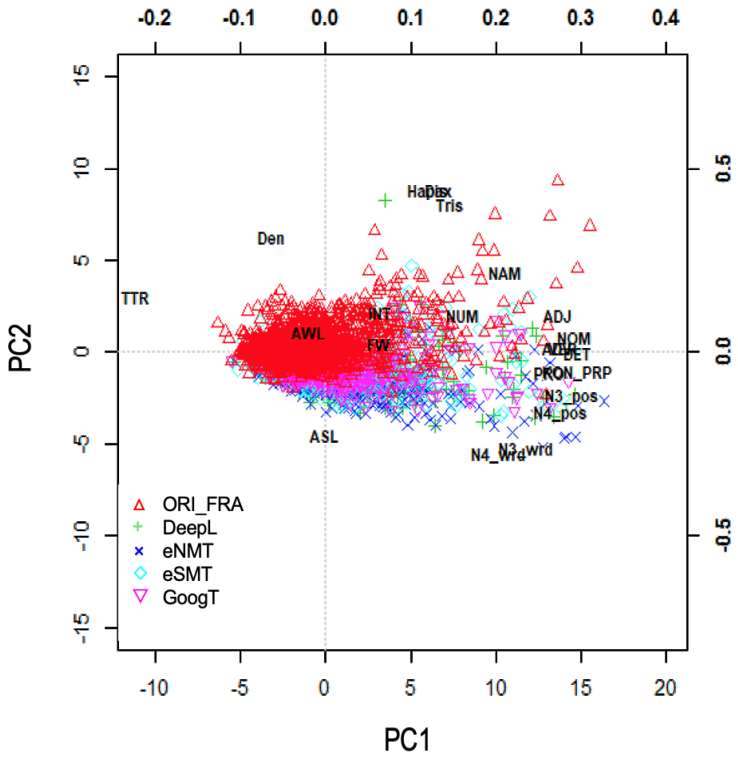

Figure 2: Same biplot as Figure 1 but without visualizing the British-English source texts.

4.2.1 Average sentence length

Figure 3 presents the ANOVA analysis for average sentence length. In this and all subsequent ANOVA graphs, interval plots are shown for every feature versus all settings, with dots representing the mean and pink lines the confidence intervals. If the intervals in different settings are far away from each other this indicates a substantial difference between the settings and if there is no overlap at all, this difference is statistically significant.

What this ANOVA reveals is that MT creates longer sentences in comparison with what is expected in French (ORI_FRA). We observe a sentence length increase of 18.2% (DeepL), 20.6% (eNMT), 19.1% (eSMT) and 15.2% (GoogleT). This is similar to human English to French translation, where a sentence length increase of around $20 - 25\%$ is considered to be normal[8].

4.2.2 ngram features

The ANOVAs of the different ngram features are depicted in Figures 4 a/b/c/d. These features are based on the top-100 most frequent combinations of 3 or 4 words (lemmas) or part-of-speech categories in both languages. Table 4 presents the top five combinations of each ngram feature in the entire French corpus.

Regarding the word ngrams both trigrams and fourgrams indicate a pronounced diffe-

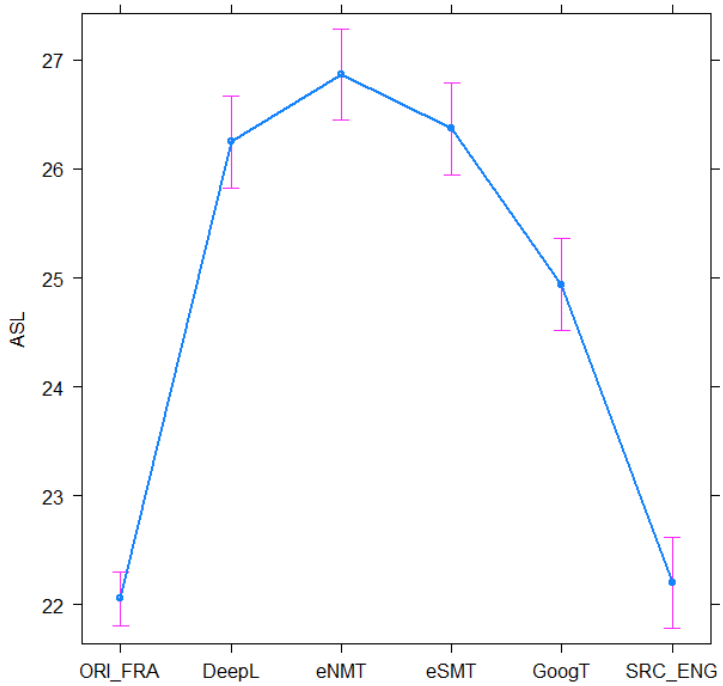

Figure 3: ANOVA average sentence length (ASL)

Table 4: Top five combinations of the French ngram features used for this study, combinations of 3 (N3_wrd) or 4 (N4_wrd) lemmatized word forms and of 3 (N3_pos) or 4 (N4_pos) part-of-speech combinations.

N3_wrd	N3_pos	N4_wrd	N4_pos
ne être pas	PRP DET NOM	*il se agir de*	NOM PRP DET NOM
il y avoir	DET NOM PRP	*Il ne y avoir*	PRP DET NOM PRP
ne avoir pas	NOM PRP NOM	*ce ne être pas*	DET NOM PRP NOM
le un du	NOM PRP DET	*se agir de un*	DET NOM PRP DET
ce être un	VER DET NOM	*avoir déclarer que le*	VER PRP DET NOM

rence between original and machine-translated French. All translation engines rely more on common words combinations and standard phrase structures. Let us have a closer look at the top three most and least frequent word ngrams in original French, compared to their position in the machine-translated texts, as presented in Table 5. The numbers represent the index of this ngram in the list of all 100 ngrams for each setting. The closer the colour is to red, the less frequent, the closer the colour is to blue, the more frequent[9]. From this table it can clearly be deduced that regarding the top three most frequent ngrams the MT engines are more or less

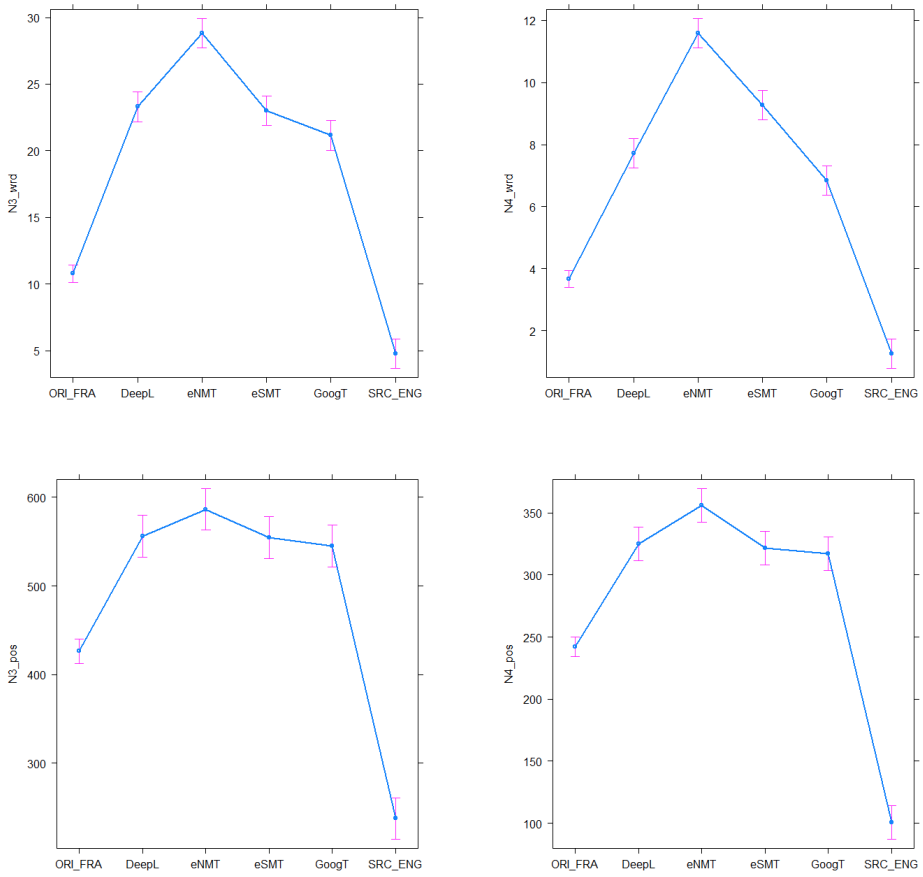

Figure 4: a/b/c/d. ANOVA analyses of the ngram features, based on combinations of 3 or 4 words (N3_wrd, N4_wrd) or part-of-speech tags (N3_pos, N4_pos).

in line with original French, especially the trigrams. However, when considering the top three least frequent ngrams, we observe that the MT engines use these ngrams much more frequently than in original French. Again this finding is more pronounced for the trigrams than for the fourgrams.

There are also quite some individual differences among the engines: for the trigrams especially eNMT uses more frequent combinations and for the fourgram features all MT engines differ significantly from each other, with the eTranslation systems standing out most clearly from the others. Inspecting the POS ngrams, there is no pronounced difference among the different MT engines for both the trigrams and fourgrams. However, machine-translated French also here clearly relies more on the same combinations of POS-tags than original French. We could say that the machine translation engines tend to "play safe". This is in line with the normalization translation universal, already found in human translations and defined by Baker

Table 5: Top three combinations of the most and least frequent lemmatized word form ngrams in the original versus the machine-translated French. The color range represents the frequency: the closer to red, the less frequent an ngram is; the closer to blue, the more frequent.

	ORI_FRA	DeepL	eNMT	eSMT	GoogT
Trigrams					
ne être pas	1	1	1	1	1
il y avoir	2	2	2	3	3
ne avoir pas	3	4	3	2	4
déclarer que il	100	74	31	32	26
déclarer que le	99	14	14	89	8
avoir déclarer que	98	3	6	13	2
Fourgrams					
il ne y avoir	1	8	3	4	6
ce ne être pas	2	5	21	12	2
il se agir de	3	2	1	3	4
avoir déclarer que il	97	19	19	15	3
du pays de Galles	96	62	72	63	36
avoir déclarer que le	95	1	4	38	1

(1993) as the exaggeration of features in the target language and conformity to its typical patterns. Given that the MT engines are being trained on large amounts of parallel human-translated data, then maybe this is why a normalization effect can also be found in MT texts. Moreover, MT engines relying more on the same word combinations also corroborates the work by Van Massenhove et al. (2019) which found that the inherent nature of data-driven MT systems to generalize over the training data has a quantitatively distinguishable negative impact on word choice, leading to less lexical diversity and bias. This is referred to as "algorithmic bias", characterized by an "exacerbation of dominant forms" (Van Massenhove et al. 2019, 223).

4.2.3 Lexical diversity

Given the findings in the previous subsection we would expect that when looking at measures of lexical diversity the MT engines reveal a similar tendency, namely of being lexically less diverse. The ANOVA plots presented in Figure 5 indeed confirm this hypothesis.

Whereas the interval plots of the original French and source English texts are similar to each other, they are more elevated than the type-token ratios present in the machine-translated texts, suggesting that machine translations are lexically less diverse. This corroborates previous similar findings by, for example, Toral (2019) or Vanmassenhove et al. (2019), which

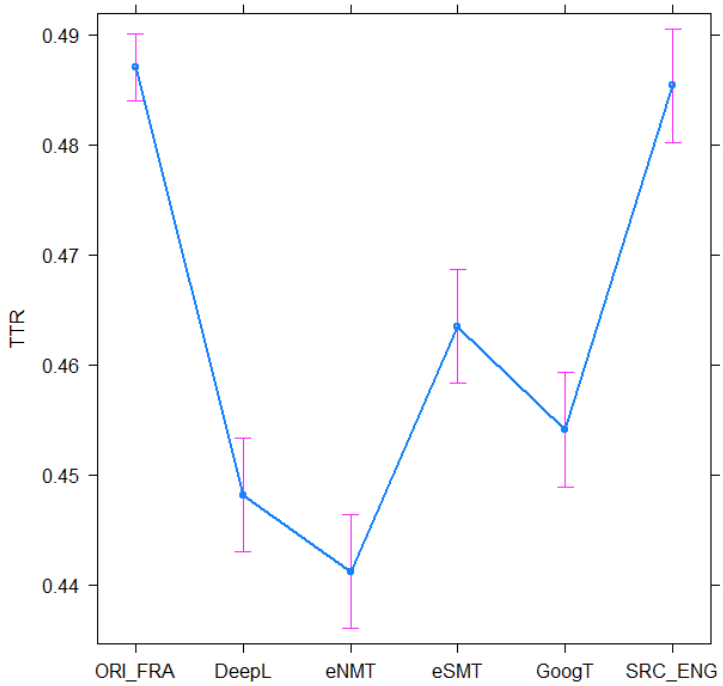

Figure 5: ANOVA analysis representing the Type-token ratio (TTR)

revealed that MT output has lower lexical diversity than human translation and that both have a lower lexical diversity than human-written, naturally composed text in the same language.

In this respect, it is also interesting to consider the legomena features, which are presented in the ANOVAs in Figures 6 a/b/c.

Overall, original French exhibits more hapax, dis and tris legomena than machine-translated French, which also hints at a difference in lexical diversity. The English source texts even comprise a much higher number of all three types of legomena than original French and one could thus expect some influence of this in the MT texts. However, this is not the case, which further underlines MT's incapability to produce lexically diverse translations independently of the lexical diversity in the source texts.

When comparing the four different MT engines we observe more or less the same tendencies for dis and tris legomena; however, the hapax legomena exhibit more pronounced differences, especially between the eTranslation systems on the one hand and DeepL and GoogleTranslate on the other hand. The eTranslation systems produce many more hapax legomena; more specifically, each text translated with the European Commission's translation engines has on average 6.25 (NMT) and 7.92 (SMT) hapax legomena per text versus 3.43 and 2.59 for DeepL and GoogleTranslate, respectively. By comparison, the average number of hapax

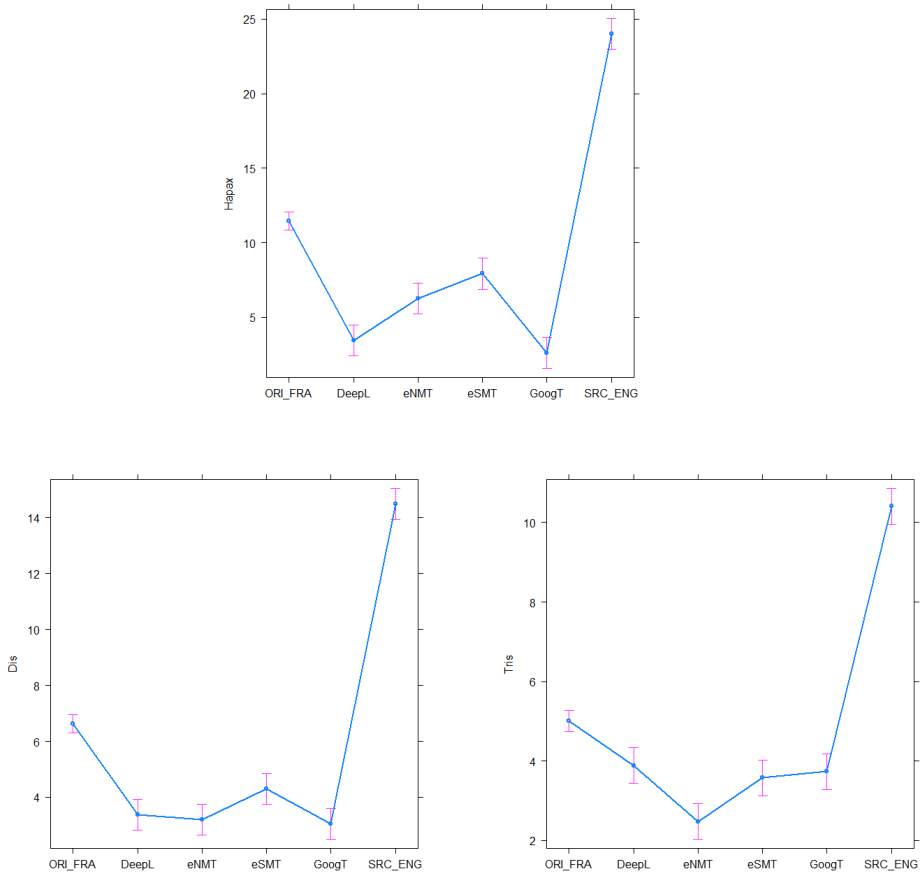

Figure 6: a/b/c. ANOVA analyses of the legomena features representing Hapax, Dis and Tris legomena

legomena in original French texts amounts to 11.52. This could imply that the output of the eTranslation systems are closer to original French when it comes to hapax legomena.

To get more insight into this, we inspected the average number of hapax legomena within each news domain (Table 6). As expected, we observe overall higher numbers for the eTranslation engines, and especially the SMT engine, for which numbers are closest to original French. However, what also draws the attention are the higher numbers – indicated in bold – in two different news domains, namely *Culture* and *Travel*, in all French corpora. This probably hints at more creative and unique language use in these two domains.

In order to get more insights into which hapax legomena are produced by the different MT engines in such domains, we manually inspected two texts, one of the *Culture* and one of the *Travel* domains, with an elevated number of hapax legomena [10]. Table 7 presents the results of this analysis, each cell containing a percentage of hapax legomena corresponding to

Table 6: Average number of hapax legomena per text in every news domain in the original French texts (ORI_FRA) compared to the machine-translated text with the four different MT engines

News domains	ORI_FRA	DeepL	eNMT	eSMT	GoogT
Business	10.09	2.08	2.5	5.25	1.67
Crime	6.49	2.05	4.05	4.79	1.69
Culture	**18.42**	**5.35**	**11.84**	**14.04**	**4.31**
Environment	11.07	2.85	3.51	5.76	2.68
Health	9.94	1.85	3.64	4.92	1.69
International News	8.5	2.25	3.52	4.96	1.85
Politics	8.75	2.44	2.24	3.7	1.54
Science & Technologies	11.19	2.79	5.75	7.91	2.61
Sport	11.63	5.87	7.24	8.56	2.24
Travel	**25.94**	**6.52**	**18.2**	**20.73**	**5.93**

one of the following categories:

- **Existing:** refers to French existing words, i.e. listed in a dictionary. Examples are *diabolisé, archétypal* or *microfissuré*.

- **Understandable:** refers to words which are not "official", but which are easy to process or understand. Examples are *zombifié, cavalage* or *vampiriquement*.

- **English:** refers to words that were not translated but merely copied from English. Examples are *avid, zombified* or *torch-lit*.

- **Made-up:** refers to made-up words which are hard to understand or words which were slightly adapted from English to French standards. Examples are *vortir, tonnamment, bienbouffeur* or *torch-éclairé*.

Table 7: Percentage of hapax legomena belonging to one of the four categories, calculated separately for every MT engine

	DeepL		eNMT		eSMT		GoogleT	
	Culture	Travel	Culture	Travel	Culture	Travel	Culture	Travel
Existing	**75.0**	**80.0**	31	28.0	24.0	32.5	**77.0**	**62.5**
Understandable	12.5	0.0	0.0	0.0	0.0	0.0	15.0	12.5
English	0.0	10.0	11.0	8.0	**72.0**	**52.5**	0.0	12.5
Made-up	12.5	10.0	**57.0**	**64.0**	3.0	15.0	8.0	12.5

The highest percentages per category are indicated in bold, and we clearly observe that both DeepL and GoogleT mostly produce unique words (hapax legomena) which exist in French. The same cannot be said for the eTranslation engines: the eNMT engine produces ma-

ny made-up words whereas the eSMT engine leaves many English words untranslated. This analysis contradicts what the numbers of Table 6 suggested: eTranslation tools and especially the eSMT engine are *not* closer to original French when it comes to hapax legomena.

Especially the made-up words are a problem to be mitigated, as research on reading comprehension of NMT nonsense words has found that this deteriorates comprehension and also leads to less confidence among readers; on the contrary, comprehension questions on words that are left untranslated are often answered more correctly (Macken et al. 2019). If we consider Table 7, especially the eNMT engine produced many nonsense words.

4.2.4 Verb, common noun, and proper name frequency

We conclude our discussion of the results by presenting one ANOVA where the French machine-translated texts seem to exhibit interference from the English source texts. Figure 7 presents the ANOVA of the frequency of the part-of-speech category verbs. Here, we observe a clear difference between original French and each of the machine-translated French corpora, which, in turn, are closer to the source English corpus.

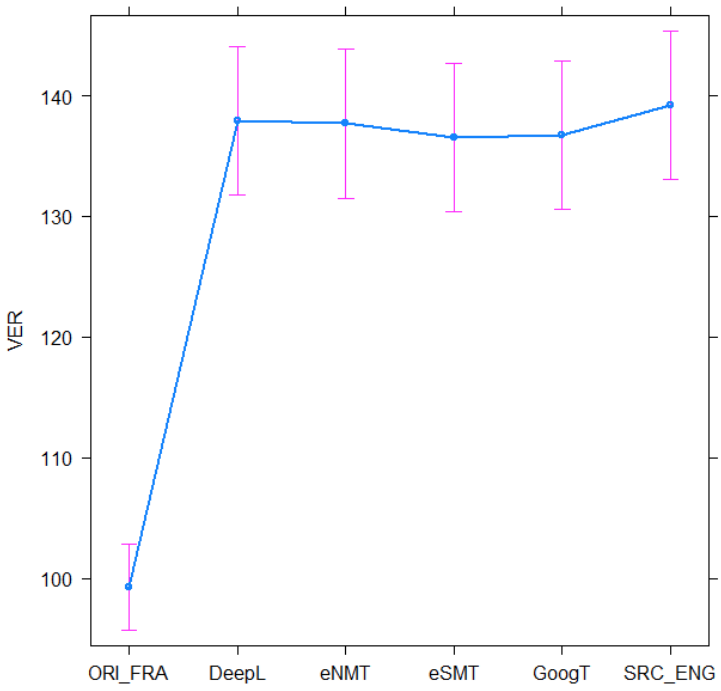

Figure 7: ANOVA analysis of the frequency of verbs (VER)

French, which is well-known to rely heavily on nominalizations, uses fewer verbs than English (compare *dans mon enfance*, ... and *when I was a child*, …). Remarkably, the corpora of machine-translated French texts all display a higher frequency of verbs than the corpus

of original French and about the same frequency as the corpus of source English texts. This hints at a shining-through effect. An increase of verbs' frequency by some 30% to over 40% compared to original French constitutes a non-negligible 'overuse'.

Given the results presented in Figure 7, one would expect a lower use of common nouns in MT French. However, as is shown by Figures 8a/b, this is not the case: common nouns (NOM) are actually (much) more frequent in MT French than in both the English source texts and original, untranslated French. Such an overuse of common nouns requires further investigation. The frequency of proper names (NAM), however, shows a highly significant decrease between the English source texts and French MT texts. Probably this is part of the explanation: proper names are somehow 'turned into' common nouns in MT. More research is clearly needed to investigate this issue.

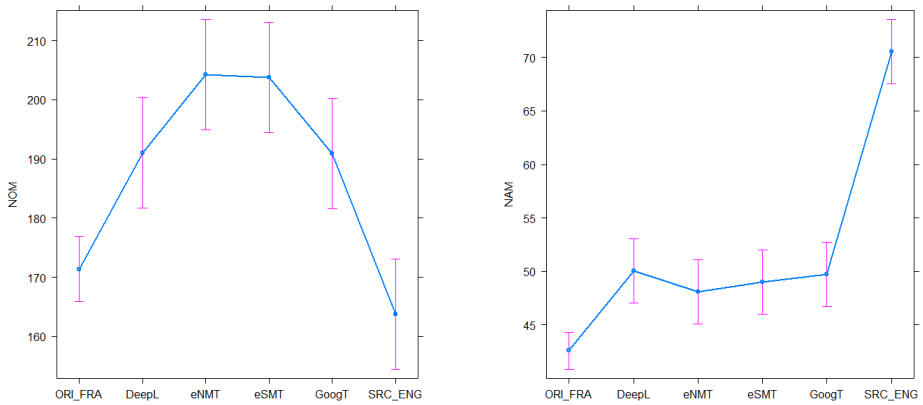

Figure 8: a/b. ANOVA analyses of the frequency of common (NOM) and proper (NAM) nouns (NOM)

5. Conclusion

In this article, we investigated the existence of machine translationese in English to French machine translations. Using the methodology and statistical techniques from corpus-based translation studies, a corpus of British English press texts was translated into French, using four different machine translation systems, and compared to French original, untranslated texts. After automatically preprocessing all texts, 22 language-independent features were extracted and subsequently the entire data matrix was analyzed with Principal Component Analysis.

This analysis revealed a distinction between original and machine-translated French, mainly due to five language-independent features: average sentence length and four features pertaining to formulaicity as expressed by combinations of three or four words or part-of-

speech combinations (ngram features). This was further explored by analysing ANOVA tests that were carried out for all features in the different settings.

Regarding sentence length, we observe a similar tendency as with human translation from English to French, namely an increase of around $20 - 25\%$. When considering the top-100 combinations of three or four words or part-of-speech tags, significant differences are found between original and machine-translated French. Similar to what was found in previous studies, the machine-translated texts thus tend to rely much more on the same word combinations, a phenomenon referred to as the "algorithmic bias" (Van Massenhove et al. 2019). Moreover, because MT systems are trained on huge amounts of human-translated parallel data this is also in line with the normalization translation universal (Baker 1993).

The ANOVA analyses also uncovered machine translationese for measures of lexical diversity. The type-token ratios of the original French are more elevated than the ones present in the machine-translated texts, corroborating previous research which found that machine translations are less lexically diverse (Toral 2019 and Vanmassenhove et al. 2019). Overall, original French exhibits more hapax, dis and tris legomena than machine-translated French, which also hints at a difference in lexical diversity. Especially the hapax legomena ANOVA yields differences among the different MT engines, suggesting that the SMT engine is closest to original French. However, upon closer inspection we discovered that this engine just leaves many words untranslated. The same analysis revealed that all MT engines also produce nonsense words, especially the eNMT engine, which is something to be avoided as this can hamper reading comprehension (Macken et al. 2019). When considering all these features, the eNMT system comes out as the one exhibiting most machine translationese and Google Translate as the one exhibiting the least.

This study has some limitations in that it only focused on original versus machine-translated French in one genre, namely press texts, and in that all features were calculated with the help of automatic preprocessing, which is not necessarily 100% accurate. Nevertheless, within a principled approach to MT tools, by both professionals and translation students who need to acquire MT literacy in order to work with the machine, such results are interesting as they provide information on what should be focused on during the post-editing process. In the case of full PE, where high quality is expected, in the light of our results, EN-FR MT output should be checked for the five linguistic features showing significant differences between machine-translated French and original French. In combination with the language-specific features investigated in Loock (2018, 2020), these independent features can provide a check-list for post-editors (e.g. reduce length of sentences), in order to try and reach linguistic homogenization with the original language, the holy grail of any translator trying to meet the invisibility demands of the high-quality end of the market.

Notes

[1] Human evaluation also has its methodological shortcomings; see Läubli et al. (2020) for an interesting discussion of which aspects should make up a human evaluation of MT output.

[2] Note that not all CBTS case studies consider the differences between original and (human) translated language to lead to translationese, a negative term suggesting that translations should be improved. Quite a number of studies actually interpret the differences as being the result of the natural influence of translation universals (simplification, normalization, explicitation, levelling out), originally defined in Baker (1993) but widely criticized since, leading rather to a "third code" for translated texts, a term that does not imply any value judgment (Gellerstam 2005, 202).

[3] https://www.deepl.com/translator

[4] www.linguee.com

[5] https://translate.google.com/

[6] https://ec.europa.eu/info/resources-partners/machine-translation-public-administrations-etranslation$_e$n

[7] We would like to thank the European Commission's Directorate-General for Translation for giving us access to eTranslation.

[8] Many translation agencies often provide tables with expected expansion rates, and the one for EN-FR translation mostly amounts to $20-25\%$, see for example https://www.versioninternationale.com/details-taux+de+foisonnement+en+traduction+-+anglais+francais+allemand-395.html

[9] Please note that "least frequent" should be taken with a grain of salt as all ngram analyses are based on the top-100 most frequent ngrams.

[10] The titles of the two texts are "Zombies: A Cultural History review – a grave injustice" (Culture domain) and "Plan your own Grand Tour of Namibia - our expert's ultimate itinerary" (Travel domain).

References

Baayen, Rolf Harald. 2008. *Analyzing Linguistic Data: A Practical Introduction to Statistics using R*. Cambridge: Cambridge University Press.

Babych, Bogdan. 2014. "Automated MT Evaluation Metrics and their Limitations." *TradumàTica: Tecnologies de la Traducció* 12: 464-470.

Baker, Mona. 1993. "Corpus Linguistics and Translation studies: Implications and Applications". In *Text and Technology: In Honour of John Sinclair*, ed. by Mona Baker, Gill Francis, and Elena Tognini-Bonelli, 223-250. Amsterdam and Philadelphia: John Benjamins.

Bentivogli, Luisa, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. 2016. "Neural versus Phrase-based Machine Translation Quality: A Case Study." In *Proceedings of Con-*

ference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, United States, 1-5 November 2016, 257-267. http://www.aclweb.org/anthology/D16-1000 (consulted 25.09.2020).

Bojar, Ondřej, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. "Findings of the 2015 Workshop on Statistical Machine Translation." In *Proceedings of the 10th Workshop on Statistical Machine Translation*, Lisbon, Portugal, 17-18 September 2015, 1-46. http://www.statmt.org/wmt15/pdf/WMT01.pdf (consulted 25.09.2020).

Bojar, Ondřej, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. "Findings of the 2016 Conference on Machine Translation." In *Proceedings of the 1st conference on machine translation*, Berlin, Germany, August 2016, 131-198. https://www.aclweb.org/anthology/W16-2301/ (consulted 25.09.2020).

Bowker, Lynne, and Jairo Buitrago Ciro. 2019. *Machine Translation and Global Research: Towards Improved Machine Translation Literacy in the Scholarly Community*. Bingley: Emerald Publishing.

Castilho, Sheila, Joss Moorkens, Federico Gaspari, Iacer Calixto, John Tinsley, and Andy Way. 2017a. "Is Neural Machine Translation the New State of the Art?" *The Prague Bulletin of Mathematical Linguistics* 108: 109-120.

Castilho, Sheila, Joss Moorkens, Federico Gaspari, Rico Sennrich, Vilelmini Sosoni, Yota Georgakopoulou, Pintu Lohar, Andy Way, Antonio-Valerio Miceli Barone, and Maria Gialama, 2017b. "A Comparative Quality Evaluation of PBSMT and NMT Using Professional Translators." In *Proceedings of the Machine Translation Summit XVI*, Nagoya, Japan, 18-22 September 2017, Vol. 1, 116-131. http://aamt.info/app-def/S-102/mtsummit/2017/conference-proceedings/ (consulted 25.09.2020).

Daems, Joke, Orphée De Clercq, and Lieve Macken. 2017. "Translationese and Post-editese : how Comparable is Comparable Quality?." *Linguistica Anteverpiensia New Series – Themes in Translation Studies* 16: 89-103.

Daems, Joke, and Lieve Macken. 2019. "Interactive Adaptive SMT vs. Interactive Adaptive NMT: A User Experience Evaluation." *Machine Translation* 33(1): 117-134.

De Sutter, Gert, Marie-Aude Lefer and Isabelle Delaere . eds. 2017. *Empirical Translation Studies: New Theoretical and Methodological Traditions*. Berlin: Mouton de Gruyter.

De Sutter, Gert, Bert Cappelle, Orphée De Clercq, Rudy Loock, and Koen Plevoets. 2017. "Towards a Corpus-based, Statistical Approach of Translation Quality: Measuring and Visualizing Linguistic Deviance in Student Translations." *Linguistica Anteverpiensia New*

Series – Themes in Translation Studies 16: 25-39.

Evert, Stefan, and Stella Neumann. 2017. "The Impact of Translation Direction on Characteristics of Translated Texts. A Multivariate Analysis for English and German." In *Empirical Translation Studies. New Theoretical and Methodological Traditions*, ed. by Gert De Sutter, Marie-Aude Lefer, and Isabelle Delaere, 47-80. Berlin: Mouton de Gruyter.

Forcada, Mikel L. 2017. "Making Sense of Neural Machine Translation." *Translation Spaces* 6(2): 291-309.

Gellerstam, Martin. 1986. "Translationese in Swedish Novels Translated from English." In *Translation Studies in Scandinavia*, ed. by Lars Wollin, and Hans Lindquist, 88-95. Lund: CWK Gleerup.

Gellerstam, Martin. 2005. "Fingerprints in Translation". In *In and Out of English: For Better, For Worse?*, ed. by Gunilla Anderman, and Margaret Rogers, 201-213. Clevedon: Multilingual Matters.

Hassan, Hany, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Junczys-Dowmunt Marcin, Huang Xuedong, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Rengian Luo, Aruk 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 (consulted 25.09.2020).

Isabelle, Pierre, Colin Cherry, and George Foster. 2017. "A Challenge Set Approach to Evaluating Machine Translation." In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, 7-11 September 2017, 2486-2496. https://www.aclweb.org/anthology/D17-1263/ (consulted 25.09.2020).

Jenset, Gard, and Barbara McGillivray. 2012. "Multivariate Analyses of Affix Productivity in Translated English." In *Quantitative Methods in Corpus-Based Translation Studies*, ed. by Michael P. Oakes, and Meng Ji, 301-324. Amsterdam and Philadelphia: John Benjamins.

Kurokawa, David, Cyril Goutte, and Pierre Isabelle. 2009. "Automatic Detection of Translated Text and its Impact on Machine Translation." In *Proceedings of Machine Translation Summit XII*, Ottawa, Canada, 10-12 July 2009, 81-88.

Lapshinova-Koltunski, Ekaterina. 2013. "VARTRA: A Comparable Corpus for Analysis of Translation Variation." Paper presented at the 6th Workshop on Building and Using Comparable Corpora, Sofia, Bulgaria, August 2013.

Lapshinova-Koltunski, Ekaterina. 2015. "Variation in Translation: Evidence from Corpora." In *New directions in Corpus-based Translation Studies*, ed. by Claudio Fantinuoli, and Federico Zanettin, 93-114. Berlin: Language Science Press.

Läubli, Samuel, Sheila Castilho, Graham Neubig, Rico Sennrich, Qinlan Shen, and Antonio Toral. 2020. "A Set of Recommendations for Assessing Human-Machine Parity in Language Translation." *Journal of Artificial Intelligence Research* 67: 653-672.

Laviosa, Sara. 2002. *Corpus-Based Translation Studies: Theory, Findings, Applications*. Amsterdam and New York: Rodopi/Leiden : Brill.

Loock, Rudy 2018. "Traduction automatique et usage linguistique: une analyse de traductions anglais-français réunies en corpus." *Meta: le journal de traducteurs/Meta: Translators' Journal* 63(3): 785-805.

Loock, Rudy. 2019. "Parce que 'grammaticalement correct' ne suffit pas: le respect de l'usage grammatical en langue cible." In *La formation grammaticale du traducteur: enjeux didactiques et traductologiques*, ed. by Michel Berré, Béatrice Costa, Adrien Kefer, Céline Letawe, Hedwig Reyter, and Gudrun Vanderbauwhede, 179-194. Villeneuve d'Ascq: Presses Universitaires du Septentrion.

Loock, Rudy. 2020. "No More Rage Against the Machine: How the Corpus-based Identification of Machine-translationese can Lead to Student Empowerment." *The Journal of Specialised Translation* 34: 150-170.

Macken, Lieve, Laura Van Brussel, and Joke Daems. 2019. "NMT's Wonderland where People Turn into Rabbits. A Study on the Comprehensibility of Newly Invented Words in NMT Output." *Computational Linguistics in the Netherlands Journal* 9: 67-80.

Macken, Lieve, Daniel Prou, and Arda Tezcan. 2020. "Quantifying the Effect of Machine Translation in a High-quality Human Translation Production Process." *Informatics* 7(12). http://hdl.handle.net/1854/LU-8660184

Moorkens, Joss, Sheila Castilho, Federico Gaspari, and Stephen Doherty. eds. 2018. *Translation Quality Assessment: From Principles to Practice*. Berlin: Springer.

Olohan, Maeve 2004. *Introducing Corpora in Translation Studies*. London and New York: Routledge.

Papineni, Kishore, Salim Roukos, Todd Ward, and Zhu Wei-Jing. 2002. "Bleu: a Method for Automatic Evaluation of Machine Translation." In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, United States, 7-12 July 2002, 311-318. https://dl.acm.org/citation.cfm?doid=1073083.1073135 (consulted 25.09.2020).

Rossi, Caroline, and Jean-Pierre Chevrot. 2019. "Uses and Perceptions of Machine Translation at the European Commission." *The Journal of Specialised Translation* 31: 177-200.

Tezcan, Arda, Joke Daems, and Lieve Macken. 2019. "When a 'Sport' is a Person and Other Issues for NMT of Novels.. In *Proceedings of the Qualities of Literary Machine Translation*, Dublin, Ireland, August 2019, 40-49. https://www.aclweb.org/anthology/W19-7306/ (consulted 25.09.2020).

Toral, Antonio. 2019. "Post-editese: An Exacerbated Translationese." In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, Dublin, Ireland, August 2019, 273-281. https://www.aclweb.org/anthology/W19-6627/ (consulted 25.09.2020).

Toral, Antonio, Sheila Castilho, Ke Hu, and Andy Way. 2018. "Attaining the Unattainable?

Reassessing Claims of Human Parity in Neural Machine Translation." Paper presented at the 3rd conference on machine translation, Brussels, Belgium, October 2018, 113-123. https://www.aclweb.org/anthology/W18-6312/ (consulted 25.09.2020).

Van Brussel, Laura, Arda Tezcan, and Lieve Macken. 2018. "A Fine-grained Error Analysis of NMT, PBMT and RBMT Output for English-to-Dutch." In Proceedings of the Eleventh International Conference on Language Resources and Evaluation, Miyazaki, Japan, 7-12 May 2018, Miyazaki, Japan, 2018, 3799-3804. https://www.aclweb.org/anthology/L18-1600.pdf (consulted 25.09.2020).

Van de Kauter, Marjan, Geert Coorman, Els Lefever, Bart Desmet, Lieve Macken, and Veronique Hoste. 2013. "LeTs Preprocess: The Multilingual LT3 Linguistic Preprocessing Toolkit." *Computational Linguistics in the Netherlands Journal* 3: 103-120.

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 Volume 1: Research Track*, Dublin, Ireland, August 2019, 222-232. https://www.aclweb.org/anthology/W19-6622/ (consulted 25.09.2020).

Yamada, Masaru. 2019. "The Impact of Google Neural Machine Translation on Post-editing by Student Translators." *The Journal of Specialised Translation* 31: 87-106.