

Loock, R. (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. <https://doi.org/10.26034/cm.jostrans.2020.141>

This article is published under a *Creative Commons Attribution 4.0 International* (CC BY):
<https://creativecommons.org/licenses/by/4.0>



© Rudy Loock, 2020

No more rage against the machine: how the corpus-based identification of machine-translationese can lead to student empowerment¹

Rudy Look, University of Lille and CNRS Research unit 'Savoirs, Textes, Langage'

ABSTRACT

The aim of this article is to show how a linguistic analysis of a corpus of machine-translated texts, both quantitative and qualitative, can empower translation trainees by helping them define their added value over machine translation (MT) systems. In particular the aim is to show that MT, even when providing grammatically correct output, does not comply with linguistic usage, thus failing to provide natural-sounding translations as expected in today's market for specialised translation. Following two avenues left open for future research in Look (2018), this article provides the results of a corpus analysis of EN-FR machine-translated texts using 3 MT systems: DeepL (NMT) and the European Commission's eTranslation in both its SMT and NMT versions. The quantitative results show that the linguistic characteristics of machine-translated texts differ from French original texts, with an almost systematic over-representation of a series of linguistic features, possibly but partially due to source language interference, while the qualitative analysis of a sample reveals finer-grained results (e.g. variability of results depending on (N)MT tool, frequency of adverb deletion). It is then explained how such results, leading to the identification of 'machine-translationese', are meant to be used in an educational setting to improve translator education, by (i) making students aware of the gap between machine-translated texts and original texts, and (ii) providing them with information on what to focus on during the post-editing process.

KEYWORDS

Machine translation, corpus analysis, translation training, language use, post-editing.

1. Introduction

It is now obvious that the technological progress of machine translation (MT) cannot be ignored by professionals in the language service industry, and as a consequence by translation trainers. Advances in MT results have been more than significant since the arrival of neural machine translation (NMT, Forcada 2017) a few years ago, which gives priority to the fluency of the target language, sometimes at the expense of fidelity to the source text, making NMT errors more difficult to identify, as has been shown by several experiments with professionals or students (e.g. Castilho *et al.* 2017a, 2017b, Yamada 2019). This is a challenge for trainee translators and for translation trainers, who need to future-proof their translation programs. Reports such as the *2018 European Language Industry Survey Report* show that for the first time, more than half of European translation companies now use MT. Students thus need to receive specific training and experiment with the use of different MT tools for various kinds of translation projects, as well as different types of post-editing (PE), the final aim being to show them how to work with the machine in a human-centred approach.

To understand what the machine can or cannot do, evaluation of MT output is crucial: trainee translators need to be able to determine when to use the technology, when it is efficient, and also what to focus on during the PE process. I agree with Moorkens (2018: 2) that “NMT output has many weaknesses as well as strengths” and believe that students should be made aware of these. In other words, it is important to define with them how human translators – sometimes called ‘biotranslators’² – can work with, and not against, the machine. This is where I believe a *linguistic* analysis of MT output can help aspiring professionals to become aware of their added value, by measuring the gap between expected norms, by which I mean linguistic usage in addition to rules, in the original language and the observed norms in machine-translated texts. Specifically, in this article, usage will focus on the *frequency* of some linguistic features in original and machine-translated French.

In Loock (2018), I conducted a linguistic analysis of a comparable corpus of EN-FR machine-translated texts and original French texts. The focus was on linguistic usage beyond grammatical correctness through the analysis of a series of linguistic features, both lexical and grammatical (e.g. use of lemmas like *thing* vs. *chose*, derived adverbs, or existential constructions, see 3.2 for a complete list). Machine-translated texts were obtained from two MT systems: DeepL, a publicly-available generic NMT tool, and eTranslation, the MT tool developed by the Directorate-General for Translation of the European Commission, which at the time (spring 2018) was still using the previous paradigm of Statistical MT (SMT). The results showed that for the features investigated, machine-translated texts showed significant differences from original French in terms of frequencies. This paper follows two avenues of future research following Loock (2018). First, now that eTranslation has become an NMT tool, it is possible to compare two NMT systems (DeepL vs. eTranslation) as well as two versions of the same tool (NMT eTranslation vs. SMT eTranslation). Second, a finer-grained, qualitative analysis of machine-translated texts is provided thanks to a sample parallel corpus. I draw here a distinction between comparable corpora, containing independent samples (e.g. original English and original French, or in this paper translated French vs. original French), and parallel corpora containing original texts in one language and their translation in at least one other language.

The article is organised as follows. I first explain the approach and the type of linguistic analysis conducted (section 2) before providing information on the corpus material and methodology (section 3). I then provide the results of the different analyses of the comparable corpora, and discuss the implications for student training (section 4). A final section is dedicated to the qualitative analysis of a parallel corpus containing a sample of NMT translated texts aligned with their original English source texts (section 5).

2. A linguistic evaluation of MT output

Ever since the early days of MT with rule-based systems (RBMT), researchers have tried to determine reliable ways of evaluating MT output so as to improve the results (see Moorkens *et al.* 2018 for a series of recent studies on the translation quality assessment of MT output). A lot of attention has been paid to automatic evaluation with the use of metrics like BLEU (BiLingual Evaluation Understudy, Papineni *et al.* 2002), METEOR (Metric for Evaluation of Translation with Explicit ORdering, Banerjee and Lavie 2005), or ROUGE (Recall-Oriented Understudy for Gisting Evaluation, Lin 2004) to name a few. Some researchers have also focused on human evaluations to try and compensate for the limitations of automatic evaluations (see e.g. Koehn 2010 or Hartley and Popescu-Belis 2004 for a discussion on the limitations of evaluation metrics). Different evaluation methods have been developed: the ranking of evaluations by professionals or non-professionals according to quality perception (e.g. Bojar *et al.* 2015); the amount of PE necessary to make a translated text acceptable (e.g. Koehn and Germann 2014, Bentivogli *et al.* 2016); error identification and classification (e.g. Federico *et al.* 2013). Some studies even combine different methods (e.g. Popović *et al.* 2013) or compare human evaluation with metrics-based evaluations (e.g. Castilho *et al.* 2017a, 2017b, Shterionov *et al.* 2018).

In addition, some researchers have been trying to set up *linguistic* evaluations of MT output, based on the analysis, quantitative and/or qualitative, of specific language features, some of them language-dependent and others language-independent such as lexical variety. For instance, Isabelle *et al.* (2017) evaluates several MT systems thanks to a series of isolated sentences with specific linguistic features (e.g. position of pronouns, presence of stranded prepositions, expression of movement) known for being problematic for EN-FR translations because of morpho-syntactic, lexico-syntactic and syntactic divergences between the two original languages. Also, some researchers have analysed machine-translated texts compiled as electronic corpora, in the same vein as what has been done in corpus-based Translation Studies (CBTS) since the 1990s and Baker's (1993) seminal paper, using the tools of corpus linguistics to uncover differences between original and translated language (see e.g. Laviosa 2002, Olohan 2004). The aim of studies conducted on collections of machine-translated texts is for instance to calculate the frequencies of some specific linguistic features in translated texts, in comparison with original language or other types of translation. For instance, Macketanz *et al.* (2017) provides a comparative analysis of three MT systems (RBMT, SMT, NMT) through the analysis of 100 segments extracted from technical documentation translated from English into German and compiled as an electronic corpus. Different linguistic features were observed: use of imperatives, compounds, question marks, particles, etc. Interestingly, their results show that overall MT systems are comparable, each having its own strengths and weaknesses. Another example is Lapshinova-Koltunski

(2015), who compares EN-DE translations for 7 registers performed with the use of different translation tools: (i) no tools at all, (ii) use of a CAT tool, (iii) use of MT (one RBMT, two SMT systems). These translations are also compared with original texts written in German. Their analysis is definitely in line with CBTS research, as the aim is to uncover what has been called 'translation universals' (defined originally in Baker 1993, and widely criticised since): simplification (through the analysis of lexical density and variety), explicitation (through the presence of explicit cohesion markers), and normalisation vs. source-language interference (through the quantification of verbs). In the same vein, in comparison with human translations and based on the analysis of *ca.* 2 million sentence pairs extracted from the Europarl corpus with language pairs being EN-FR and EN-ES, Vanmassenhove *et al.* (2019) have shown that machine-translated texts fail to reach the lexical richness found in human translations (itself lower than in original texts). According to the authors, this is due to an overuse by the MT systems of more frequent words and an underuse of less frequent words because of "a form of algorithmic bias". Finally, in Look (2018), I have compared EN-FR machine-translated texts with original texts written in French. Using two different MT systems, one generic (<https://www.deepl.com>) and one specific to an international organisation (the European Commission's eTranslation tool), I observed in two corpora of EN-FR machine-translated texts the frequencies of a series of linguistic features, lexical and syntactic (see section 3.2 for a list), in comparison with texts written originally in French. The results show that machine-translated texts significantly diverge from the norms of original French, with systematic over-representations of the observed linguistic features. However, a comparison with the frequencies in the English source texts shows that source language interference cannot be the only explanation, hence the need for a qualitative analysis.

What these studies point to is the existence of 'machine translationese'/'MTese' for raw MT output, while other studies also uncover 'post-editedese' for MTPE (see Daems *et al.* 2017 and Toral 2019 for a discussion and contradictory results), alongside translationese, to be found in human translations as has been shown in corpus-based Translation Studies since the mid-1990s. Such linguistic evaluation seems to be crucial now that NMT systems are in production use: a lot of progress has undeniably been made, which makes post-editing a more difficult task than with SMT systems (see Castilho *et al.* 2017a, 2017b and Yamada 2019 for experiments with professionals and students respectively: if PE is faster with NMT, the errors to be edited are more 'human-like' and thus more difficult to identify, in particular for students). As the NMT systems tend towards target language fluency, sometimes at the expense of fidelity to the source text, not only are accuracy errors more difficult to identify, but also the grammatical correctness of the MT output might give the illusion that the translation takes into account the usage-based norms of the target language. However, for high quality translation, grammatical correctness does not suffice: naturalness and idiomaticity are expected. Another

important aspect is also the need to define how translators can work with MT tools without fear of losing control (see Rossi and Chevrot 2019 for example on how MT is perceived at the DGT), and I believe that a linguistic analysis of MT output can help as a sensitisation tool. Such evaluation can help translation students “demystify” (Moorkens 2018: 2) MT output and the disruptive technology in general, at a time when translation trainers are currently figuring out ways of teaching MT (see e.g. Massey and Ehrensberger-Dow 2017, Rossi 2017, Moorkens 2018, Faria Pires 2018, Guerberof Arenas and Moorkens 2019, Martikainen and Mestivier 2019). This paper aims to contribute to the debate.

3. Materials and methodology

3.1. Corpus material

The comparable corpus consists of two main corpora: (i) texts written in original French and (ii) texts translated from English into French thanks to three different MT tools: (iia) DeepL (NMT), (iib) eTranslation SMT, and (iic) eTranslation NMT.

DeepL is an NMT tool freely available online for everyone to use, trained on the corpus used by Linguee (www.linguee.com), and known for its sometimes impressive target language fluency, explaining why DeepL is a particularly relevant MT tool for the linguistic analysis conducted here. English texts were copied and pasted into the source text window, with a limit of 5,000 words (longer texts were therefore divided into different parts). As for eTranslation, developed by the European Commission’s DGT, it has restricted access and is not available to the public³. For the EN-FR language pair, the tool was an SMT tool (then called MT@EC) until September 2018, when the NMT version was launched. It is not a generic tool, since it is trained on institutional texts for internal use. The translations, obtained thanks to the uploading of the texts onto the platform with a limit of 50 files a day, were collected in March-April 2018 for DeepL and eTranslation SMT, and in December 2018 for eTranslation NMT.

The original French texts and the original English texts that were translated automatically were extracted from the TSM press corpus⁴ (Loock 2019), which is currently (early 2020) a 2-million-word corpus of press texts extracted from the British press, the American press, and the French press for a series of topics: business and finance, crime, culture, environment, health, etc. At the time the corpus study was conducted (from spring to winter 2018), the corpus contained 1.2 million words for 1,094 French texts and 927 English texts (437 for US English, 490 for UK English). For the current experiments, all the original French texts were selected for the study, but for EN-FR machine-translated texts, only the British sub-corpus was selected and submitted to the three MT tools mentioned above. Table 1 provides a description of the content of the part of the TSM press corpus that was used for the present study and Table 2 details the corpus that was used for the linguistic analysis.

	Original UK English	Original French
Economy and Finance	6136	36,964
Crime	43,710	93,347
Culture	46,839	78,897
Environment	32,367	88,574
Health	28,170	65,024
International News	29,168	65,354
Politics	46,901	98,540
Science and Technologies	47,213	97,252
Sports	43,766	97,367
Travel	50,056	97,351
Total number of words	374,326	818,670
Number of texts	490	1,094

Table 1. Content of the TSM press corpus (2018) used for the present study

	Original French	EN-FR translations with DeepL (NMT)	EN-FR translations with eTranslation (SMT)	EN-FR translations with eTranslation (NMT)
Number of texts	1094	490	490	490
Number of words	816,338	442,439	445,914	451,704

Table 2. The corpus used for the study

At this stage it is important to say a few words about the type of data selected for the study. The texts all belong to the press genre. However, neither DeepL nor eTranslation are trained for the translation of press texts: DeepL is meant to be a generic tool, while eTranslation is trained for institutional texts. This means that neither tool is totally fit-for-purpose, which needs to be acknowledged. Press texts were selected as the linguistic characteristics of this genre are generally not too 'specialised' (texts come from the daily quality press in France and the United Kingdom) and the vocabulary is quite general. As it was impossible for me to develop my own fit-for-purpose MT tools, I have tried to find a compromise, although this

remains a limit for the corpus study that the reader must definitely be aware of.

It is also important to mention here that the aim of this study is to compare MT output with original texts, not with other types of translated texts, in particular human translations, which is the next step in the project (see conclusion).

3.2. Methodology

All texts extracted from the TSM press corpus and all translations obtained from the three MT tools were saved as .txt files with UTF-8 encoding for analysis with an offline concordancer, namely AntConc version 3.5.7 (Anthony 2018). Part-of-Speech tagging was performed thanks to TreeTagger (Schmid 1994) for French and English using the TagAnt software version 1.2.0 (Anthony 2015). Automatic searches were performed thanks to the concordancer, with manual weeding out when necessary to remove noisy results, i.e. false positives, in particular for existential constructions, where the strings *il y+AVOIR* or *there+BE* can occur in examples that are not existential constructions, one of the features that were investigated, as illustrated by (1a/b).

(1) a. ***Il y a*** *passé beaucoup de temps.*

'He there spent a lot of time'

He spent a lot of time there.

(1) b. *The man over **there was** wearing red shoes.*

Once the raw frequencies were collected, they were normalised into frequencies per million words (pmw) to allow for comparisons between the different sub-corpora. Finally, a statistical test was used to determine whether the observed differences between original and machine-translated texts were significant or not, for each of the linguistic features: the difference between two independent proportions, where a z-ratio is calculated to measure the extent of the difference between frequencies in two independent samples, as well as a p-value (Cappelle and Loock 2013). The p-value retained to reject the null hypothesis is $p=0.01$.

The linguistic analysis focused on linguistic features which are known to be problematic for EN-FR translators, due to significant differences in frequencies between original English and original French. Because of such differences, these linguistic features can show a different frequency in translated texts because of source language interference. Many EN-FR translation or comparative grammar textbooks (e.g. Vinay and Darbelnet 1995, Chuquet and Paillard 1987, Guillemin Flescher 1986) mention such linguistic features and provide suggestions to avoid systematic, overly-literal translations leading to unnaturalness. In particular the focus here is on linguistic features for which a higher, often much higher, frequency exists in original English as opposed to original French:

- the lemma *chose* (vs. its direct equivalent *thing*);

- the lemma *dire* (vs. its direct equivalent *say*);
- the coordinator *et* (vs. its direct equivalent *and*);
- the preposition *avec* (vs. its direct equivalent *with*);
- derived adverbs ending in *-ment* (vs. their direct equivalents in *-ly*);
- existential constructions (*il y+AVOIR*) (vs. their direct equivalents *there+BE* constructions)⁵.

4. Results

4.1. Quantitative analysis

In line with the results in Loock (2018), the supplementary results for eTranslation NMT show that, almost on a systematic basis (one exception), the observed linguistic features show a significant over-representation in machine-translated French. For example, the coordinator *et* ('and') is systematically over-represented in machine-translated French, with a frequency of 18,079.52 occurrences pmw in original French and frequencies of 21,435.72 (DeepL), 21,129.63 (eTranslation SMT), and 20,980.55 (eTranslation NMT) occurrences pmw in machine-translated French (z-ratio=-13,081; p<.0001; z-ratio=-11,349; p<.0001; z-ratio=-11.425 p<.0001, respectively). Another example of a systematic over-representation but with diverging results for the different MT tools is the use of the verb *dire* ('say'): its frequency (all inflections) in original French press texts is 946.91 occurrences pmw, but is 3315.71, 1170.63 and 1157.83 occurrences pmw for French texts translated with DeepL, eTranslation SMT, and eTranslation NMT, respectively. Interestingly, these results show a very important difference between the MT systems, with DeepL showing a frequency that is 3.5 times that of original French, while with eTranslation, whether SMT or NMT, the ratio is only 1.2.

Only one feature shows no significant difference between original French and machine-translated French, and for one of the three MT tools only (eTranslation NMT): the preposition *avec*, the direct equivalent of the preposition *with*, shows no statistical difference in frequency: 3689.65 occurrences pmw for original French vs. 3734.75 for machine-translated French (z-ratio=-0.4; p-value =0.3446).

Table 3 below provides the normalised frequencies (pmw) for each of the linguistic features and the three MT tools that were used for the experiment, with the raw frequencies in brackets.

Figure 1 provides a visual summary of these results.

	Original French	DeepL (NMT)	eTranslation (SMT)	eTranslation (NMT)
CHOSE	248.67 (203)	556.01 (246)	410.39 (183)	411.77 (186)

DIRE	946.91 (773)	3315.71 (1,467)	1170.63 (522)	1157.84 (523)
ET	18,079.52 (14,759)	21,435.72 (9484)	21,129.63 (9422)	20,980.55 (9477)
AVEC	3689.65 (3012)	4289.86 (1898)	4184.66 (1866)	3734.75 (1687)*
IL Y A	850.14 (694)	1573.10 (696)	1000.19 (446)	1385.86 (626)
-ment adverbs	6978.73 (5697)	8148.02 (3605)	9071.26 (4045)	7704.16 (3480)

Table 3. Normalised (pmw) and raw frequencies for original French and machine-translated French with the different MT tools (* = difference with original French is statistically non-significant).

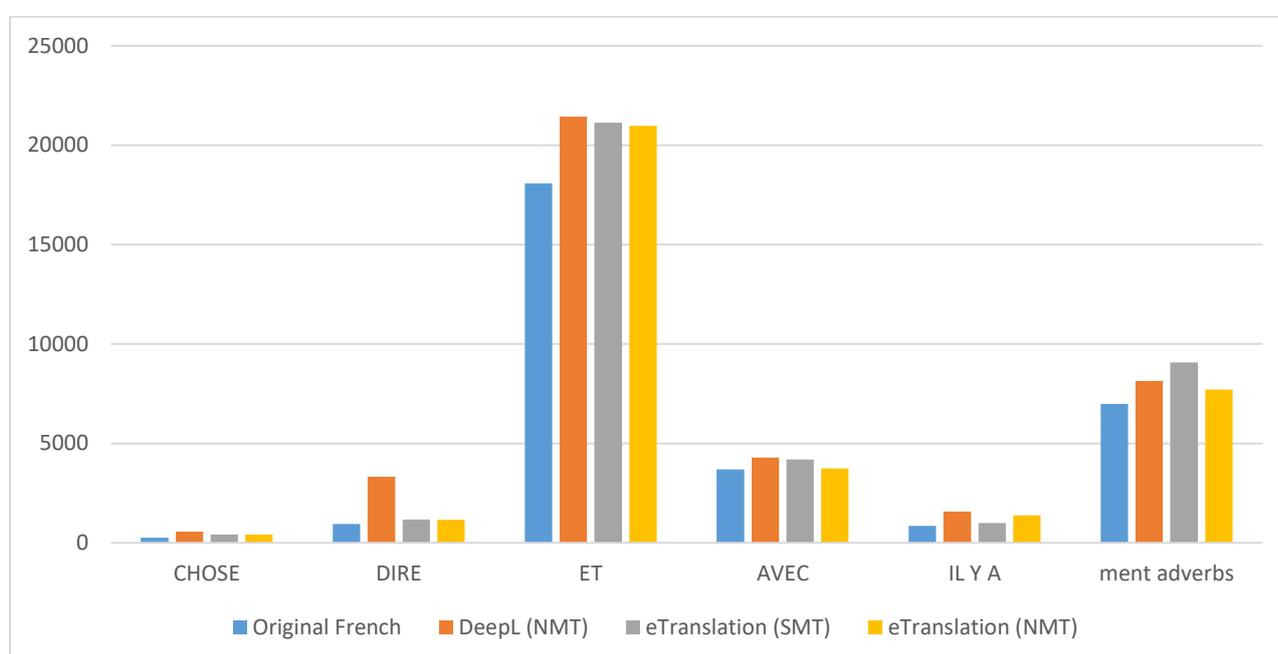


Figure 1. Normalised frequencies (pmw) of the different linguistic features in original French and EN-FR machine-translated texts

4.2. Discussion

What these quantitative results show is that in spite of significant progress, MT tools do not take into account linguistic usage in addition to grammatical rules: if grammatically-correct translations seem to be on their way, naturalness and idiomaticity still need to be improved. As in Loock (2018), the results show that even a high-quality, cutting-edge neural tool like eTranslation NMT cannot produce output showing linguistic homogenisation with original language. Depending on the purpose of the translation project, this could lead to a quality issue, and the post-editing process should take these deviations into account to try and remedy them. MT is being deployed in many contexts and in many different ways, with different expectations in terms of quality (see Way 2018 for an interesting discussion) and as a

consequence different types of post-editing: none, light or full. This is therefore a time when it has become important to measure the productivity and quality gains when using new NMT tools, in the professional world but also with students (e.g. Jia *et al.* 2019), as well as the cognitive effort necessary for PE (e.g. Koglin and Cunha 2019). It is now widely acknowledged that skills necessary for PE differ from those necessary for translation and that students need specific training to be competent post-editors (see e.g. Sycz-Opoń and Gałuskińska 2017 or Martikainen and Mestivier 2019 for experiments). The results presented here can be exploited in the case of full post-editing, “whereby the automatic translation is corrected and improved to match the quality achieved by human translation” (Screen 2019: 135). For high quality translations, being aware of specific deviations with original language can help reduce the gap with original target language norms/usage and move towards full invisibility, in the same way as being aware of gender-related errors in MT output (Vanmassenhove *et al.* 2018) should lead post-editors to focus their attention on such an issue.

Such results and the approach adopted here can also serve as a sensitisation tool to make students aware of the limits of and issues with MT systems, which seems particularly relevant when claims are being made that human parity has been reached, as was done in a Microsoft research paper (Hassan *et al.* 2018) and as is done regularly by MT developers. Becoming aware of differences between machine-translated texts and original language can help students ‘demystify’ (see section 2) the MT output which can sometimes give the illusion of perfect fluency. For example, it is possible to provide students with data showing the differences between original texts and machine-translated texts (with or without post-editing) for a series of linguistic features. Complementarily, data can be provided for human-translated texts. Students will then become aware of the existence of MTese (and even post-editease) in addition to translationese, which is characteristic of human-translated texts and of which students are generally more aware. With such data in mind, students can be asked to post-edit machine-translated texts to try and reduce the gap existing with human-translated or even original texts, with a specific focus on the results provided by corpus-based studies on MT texts, e.g. the lower lexical variety discussed in Vanmassenhove *et al.* (2019), or the over-representation of the linguistic features discussed in this paper. A complementary exercise particularly relevant here is to ask students to translate sentences or revise translations *not* using certain linguistic features, in order to develop their creativity by forcing them to steer away from literal translations (see methodology presented in Loock 2019 for a comparative grammar class for translators). Finally, such results are a good opportunity for a discussion with students on the importance of the data to be found in the corpora used to train the MT systems: if as discussed in Loock (2018) source language interference cannot be the only explanation for the observed over-representations (see also section 5), students need to be sensitised to the fact that the translated data used to train the MT

systems might show certain translationese features that get reproduced in the MT outputs. This kind of sensitisation is important, as the issue is also valid for translation memories used in CAT tools.

Ultimately, the aim is to empower students by helping them distance themselves from a disruptive technology for which they often get false ideas leading to feelings of worry. Such an approach belongs to a metacognitive approach, where students “reflect on the deployment of language technologies, by learning about the capabilities and limitations of the machines and tools with which they are and will be working” (Massey and Ehrensberger-Dow 2017: 307). It also aims to develop students’ ‘MT literacy’, a very relevant concept developed by Bowker and Ciro (2019).

If in order to develop such critical thinking, quantitative results such as those discussed above can help students become aware of the gap between original and machine-translated language, for the picture to be complete, a qualitative analysis of source texts and MT output (parallel corpus) can also be relevant so as to observe the kinds of output provided by MT systems for specific language features. This is what I now turn to in the next section.

5. A supplementary, qualitative analysis of a parallel corpus

In Loock (2018), I suggested that source language interference might explain the observed data: because of significant differences between the two original languages, a transfer occurs and the machine-translated texts show an over-representation as MT systems would perform literal translations more often than what is required for natural-sounding translations. However, the analysis showed that this was not the case and that source language interference can only be *part* of the explanation. To take one example, the frequency of *il y+AVOIR* constructions in machine-translated French cannot be explained only by the presence of *there+BE* constructions in English in the original texts. Otherwise one could expect a lack of differences between the original texts written in English and their translations into French as far as the frequencies of existential constructions are concerned. This is not the case: as shown in Figure 2, existential constructions in the English source texts have a normalised frequency of 1771.18 occurrences pmw, while in machine-translated texts, the frequencies of *il y+AVOIR* constructions range from 1000.19 to 1573.10 occurrences pmw (in original French the frequency is 850.14 occurrences pmw). These differences between original English texts and their translations are statistically significant for eTranslation (SMT and NMT), but not for DeepL (z-ratio=2.189; p-value=0.0143) (see Loock 2018 for results on other linguistic features, most of which also show diverging frequencies between original English and machine-translated French.)

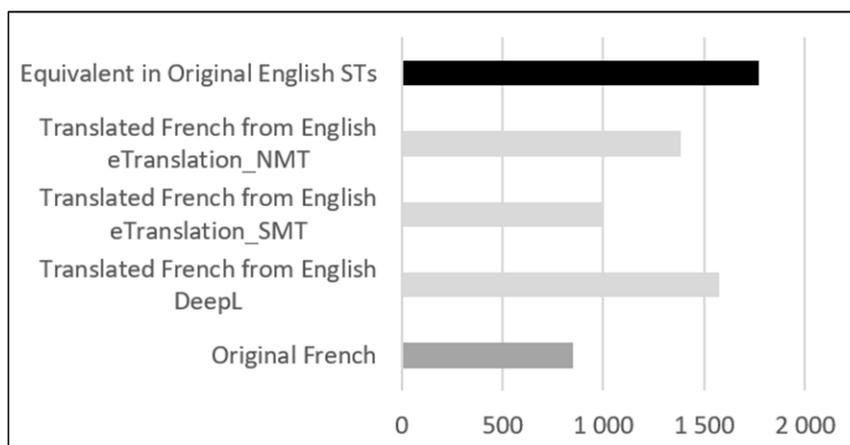


Figure 2. Normalised frequencies (pmw) of *there+BE* constructions in English and *il y+AVOIR* constructions in French in the different sub-corpora

Such differences suggest that in *some* cases at least, *there+BE* constructions are not translated with *il y+AVOIR* constructions, although from a purely syntactic point of view, this is always grammatically possible. This could be the result of a ‘stylistic’ shift, based on the data used to train the MT system, which can itself show translationese features characteristic of human-translated texts (see above), or could also be the result of statistical bias leading to the “exacerbation of dominant forms” through overgeneralisation (Vanmassenhove *et al.* 2019: 223).

In particular in an educational context, a finer-grained approach with a parallel corpus thus seems necessary for students to take a closer look at how the investigated linguistic features are actually translated, as the data cannot only be explained by simple transfer.

To conduct this qualitative analysis, a sample from three sub-corpora in the corpus was extracted: (i) original English texts, (ii) DeepL translations, (iii) eTranslation NMT translations. I extracted the first ten texts for the four first topics in the TSM press corpus (Business and Finance, Crime, Culture, Environment), for a total of 40 texts and 25,739 words. The sub-corpora were aligned at the level of the sentence thanks to the alignment tool in *Wordfast Anywhere*, with manual corrections when necessary – which was the case much more for eTranslation than DeepL output. Table 4 below provides a description of the parallel corpus, which contains 2386 aligned sentences, that is two series of 1193 sentences (original English/DeepL translation, original English/eTranslation translations). This sample is not meant to be representative but to be used as a sensitisation tool for students.

	Files	Original English	MT French DeepL	MT French eTranslation NMT	Aligned sentences
Business and Finance	10	4836	5894	6026	201
Crime	10	6941	7757	8441	344
Culture	10	6794	7730	7755	331
Environment	10	7168	8795	8936	317
TOTAL	40	25,739	30,176	31,158	1193

Table 4. Number of words and aligned sentences in the parallel corpus

Three linguistic features were selected for the manual analysis: the translation of the lemma *thing*, the translation of *-ly* adverbs, and the translation of existential *there+BE* constructions. The features were retrieved in each sentence in the original English sub-corpus, and the translations manually categorised based on the type of translations. For the translation of *thing*, translations were either a literal translation (*chose*) or an alternative noun (*élément, situation*). For *-ly* adverbs, translations were divided into literal translation (*-ment* adverb), other type of adverb (*tôt, surtout*), locution (series of words used as an adverb like *en particulier* or *de plus en plus*), prepositional phrases (e.g. *avec succès, avec force*), change of category/recategorisation (adjective). For existential constructions, a distinction was made between the direct equivalent *il y+AVOIR* construction, presentational constructions (*il existe*), and other impersonal constructions (use of impersonal *on* for example). For the three features were also added deletion (the feature was not translated/omitted but the sentence was translated), non-translations (the sentence or part of the sentence was not translated), and nonsense when the sentence could not be analysed. Table 5 below provides the detailed results of the analysis.

<i>thing</i>	DeepL	eTranslation	<i>-ly</i> adverbs	DeepL	eTranslation	Existential constructions	DeepL	eTranslation
<i>chose</i>	13	12	<i>-ment</i> Adv	138	124	<i>il y a</i>	33	25
Other N	1	0	Other type of Adv	27	23	<i>il existe</i>	0	3
Deletion	0	2	PP	19	24	Impersonal construction	0	2
TOTAL	14	14	Recategorisation	18	13	Other	2	2
			Locution	28	28	Deletion		3
			Other	13	15	NT	1	1
			Deletion	9	25	TOTAL	36	36
			Non-Translation	4	3			
			Nonsense	0	1			
			TOTAL	256	256			

Table 5. Results of parallel corpus analysis.

A first valuable piece of information is that if the number of non-translations is low, the number of deletions is quite high, in particular for the translation of *-ly* adverbs (3.5% with DeepL and about 10% with eTranslation NMT). This means that the translation of adverbs is a feature that should be checked carefully during the post-editing process, as the non-translation of an adverb can seriously impact the meaning of the sentence as in examples (2), where some information is clearly missing (note that the MT outputs provided in this section have not been post-edited in any way; the literal back-translations are only meant to provide non French-speaking readers with the content of the MT output):

(2) a. *In his address he **baldly** told his overwhelmingly climate sceptic opponents that "no challenge poses a greater threat to future generations than climate change."*
 Dans son allocution, il a déclaré à ses adversaires, qui sont majoritairement sceptiques face aux changements climatiques, qu'"aucun défi ne constitue une plus grande menace pour les générations futures que le changement climatique." (DeepL)
 Literal back-translation: In his address he told his opponents, who are overwhelmingly skeptical in relation to climate change, that "no challenge represents a bigger threat for the future generations than climate change."

(2) b. *Bringing visibility to a new image of ageing... will **hopefully** help change our attitudes towards growing older.*
 Donner de la visibilité à une nouvelle image du vieillissement... va nous aider à changer nos attitudes envers le vieillissement. (DeepL)
 Literal back-translation: Bringing visibility to a new image of ageing... will help us change our attitudes to ageing.

It was also noticed that the translation of *-ly* adverbs could lead to very severe adequacy problems, as in example (3) below, which was not the case for the translation of *thing* or existential constructions. In (3a), mature women are described as being emotional; in (3b) really good scientists become real scientists in the translation.

(3) a. *And though mature women dressing up for fun may occasionally be ridiculed, on the whole they are viewed **affectionately** (...).*
 Et si, dans l'ensemble, les femmes mûries peuvent parfois s'habiller de manière ridicule, dans l'ensemble, elles sont considérées comme affectives (...) (eTranslation NMT)
 Literal back-translation: And if, on the whole, matured women can sometimes dress in a ridiculous way, on the whole, they are considered as emotional.

(3) b. *But you know what – I know a lot of **really** good scientists at Nasa and NOAA (the National Oceanic and Atmospheric Administration), and at our universities.*
 Mais vous savez ce que je connais beaucoup de véritables scientifiques de Nasa et de NOAA (l'administration nationale des océans et de l'atmosphère), et dans nos universités. (eTranslation NMT)
 Literal back-translation: But you know that I know a lot of real scientists at Nasa and NOAA (the national administration of oceans and the atmosphere), and in our universities.

A second interesting result is the variability of the proportion of literal translations depending on the linguistic feature and the MT tool: in the case of *thing*, even if the sample is too small to be representative (only 14

occurrences in the source texts), between 85 and 92% of occurrences are translated with the direct equivalent *chose*. For existential constructions, DeepL provides a literal translation with an *il y+AVOIR* construction in 91% of cases; with eTranslation this is only 69% of cases, with quite a variety of translations, some of them quite natural-sounding (two examples are provided in (4)). For *-ly* adverbs, the results are quite different, due to the fact that some *-ly* adverbs are lexical gaps and some others have a direct equivalent but not ending in *-ment*: only 54% and 48% of *-ly* adverbs are translated with a *-ment* adverb.

(4) a. (...) *there is a danger that the UK, with its restrictive planning regulations for renewables, will find itself increasingly swimming against the global tide.*

(...) *il existe un risque que le Royaume-Uni, avec ses réglementations restrictives en matière d'aménagement du territoire pour les énergies renouvelables, s'attache de plus en plus à la marée mondiale.* (eTranslation NMT)

Literal back-translation: (...) *there exists a risk that the United Kingdom, with its restrictive regulations in terms of land management for renewable energies, gets more and more attached to the global tide.*

(4) b. *"In 2011 in Australia, we just got out of the drought, and then there is a forecast of a La Niña," he says.*

«En 2011, en Australie, nous venons de sortir de la sécheresse, puis on prévoit un «La Niña», dit-il. (eTranslation NMT)

Literal back-translation: *"In 2011, in Australia, we just come out of the drought, then we forecast a «La Niña», says he.*

Finally, it is interesting to observe some translations which clearly show that MT output can be very good and serve as inspiration for students, meaning MT should be considered a translation tool to help them in the same way translation memories and specialised electronic corpora do. Such examples are provided in (5).

(5) a. *Bringing visibility to a new image of ageing... will **hopefully** help change our attitudes towards growing older.*

Il est à espérer que la visibilité d'une nouvelle image du vieillissement... contribuera à modifier notre attitude à l'égard du vieillissement. (eTranslation NMT)

Literal back-translation: *It is to hope that the visibility of a new image of ageing... will contribute to modify our attitude to ageing.*

(5) b. (...) ***there is** no evidence to suggest that the level of violence has changed in children's films since Snow White in 1937.*

(...) *rien n'indique que le niveau de violence a changé dans les films pour enfants depuis Snow White en 1937.* (eTranslation NMT)

Literal back-translation: *nothing indicates that the level of violence has changed in movies for children since Snow White in 1937.*

(5) c. *They're portable, accessible, **constantly** improving and reworking the way we can shoot, edit and print images with minimal hardware and software.*

Ils sont portables, accessibles, en constante amélioration et retravaillent la façon dont nous pouvons filmer, éditer et imprimer des images avec un minimum de matériel et de logiciels. (DeepL)

Literal back-translation: *They are portable, accessible, in constant improvement and reworking the way we can film, edit and print images with a minimum of hardware and software.*

In addition to the results provided by a quantitative analysis of comparable data, a qualitative analysis, even though of a small sample, brings complementary, valuable information on what to focus one's attention on during the post-editing process and can help students understand what MT systems can and cannot do. In a translation master's programme, it seems very difficult to teach students how NMT actually works as most students do not have in-depth NLP (Natural Language Processing) knowledge. Although they can be sensitised to the necessity of always checking whether the data on which an MT system has been trained are fit-for-purpose (domain-specific and up-to-date), and to the ways MT can be integrated into the translation workflow, they cannot be properly trained on the technology underlying MT tools, which would require at least specific training in NLP. This means that the critical observation of MT output is crucial for them to understand when to use MT, how different systems perform, and what they should pay attention to during the post-editing process.

6. Conclusion

In this article I have suggested that the linguistic analysis of machine-translated texts compiled as electronic corpora could provide relevant information on the quality of MT output and on the kinds of elements that should require special attention during the post-editing process. In an educational setting, by focusing on specific linguistic features, translators-to-be can be sensitised to the performance and limits of MT systems and therefore define their added value over the machine, since MT output, in spite of indisputable progress, does not seem to take into account language norms such as frequencies of use. Students can then become aware of the gap that exists between original and machine-translated language, while in order to reach the invisibility required by the industry this gap should be at least reduced. To complement such observations, it would be interesting to compare machine-translated texts with human-translated texts, as translationese, or 'third code' (Frawley 1984) to use a more neutral term, is known to be a reality for human translations: total linguistic homogenisation is rarely achieved, as has been shown by numerous studies in the corpus-based Translation Studies field. This requires the compilation of a corpus of EN-FR human-translated press texts, and is left for future research.

References

- **Anthony, Laurence** (2015). *TagAnt* (Version 1.2.0). Tokyo, Japan, Waseda University. <http://www.laurenceanthony.net/software> (consulted 18.06.2019).
- **Anthony, Laurence** (2018). *AntConc* (Version 3.5.7) Tokyo, Japan, Waseda University. <http://www.laurenceanthony.net/software> (consulted 18.06.2019).

- **Baker, Mona** (1993). "Corpus linguistics and translation studies: Implications and applications." Mona Baker, Gill Francis and Elena Tognini-Bonelli (eds) (1993). *Text and technology: In Honour of John Sinclair*. Amsterdam/Philadelphia: John Benjamins, 223-250.
- **Banerjee, Satanjeev and Alon Lavie** (2005). "METEOR: An automatic metric for MT Evaluation with improved correlation with human judgments." *Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, 43rd Annual Meeting of the Association of Computational Linguistics (ACL-2005), Ann Arbor, Michigan, United States, June 2005, 65-72. <https://www.aclweb.org/anthology/W05-0909/> (consulted 22.10.2019).
- **Bentivogli, Luisa et al.** (2016). "Neural versus Phrase-Based Machine Translation quality: a case study." *Proceedings of Conference 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 18.06.2019).
- **Bojar, Ondřej et al.** (2015). "Findings of the 2015 workshop on Statistical Machine Translation." *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 18.06.2019).
- **Bowker, Lynne and Jairo Buitrago Ciro** (2019). *Machine Translation and Global Research: Towards Improved Machine Translation Literacy in the Scholarly Community*. Bingley: Emerald Publishing.
- **Cappelle, Bert and Rudy Loock** (2013). "Is there interference of usage constraints? A frequency study of existential *there* is and its French equivalent *il y a* in translated vs. non-translated texts." *Target* 25(2): 252-275.
- **Castilho, Sheila et al.** (2017a). "A Comparative quality evaluation of PBSMT and NMT using professional translators." *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 27.03.2020)
- **Castilho, Sheila et al.** (2017b). "Is Neural Machine Translation the New State of the Art?" *The Prague Bulletin of Mathematical Linguistics* 108(1), 109-120.
- **Chuquet, Hélène and Michel Paillard** (1987). *Approche linguistique des problèmes de traduction anglais-français*. Paris: Ophrys.
- **Daems, Joke, De Clercq, Orphée and Lieve Macken** (2017). "Translationese and post-editeese: How comparable is comparable quality?" *Linguistica Antverpiensia, New Series: Themes in Translation Studies* 16, 89-103.
- **Faria Pires, Loïc de** (2018). "Intégration de la traduction automatique neuronale à la formation universitaire des futurs traducteurs : pistes d'exploration." *Myriades* 4, 53-65.
- **Federico, Marcello et al.** (2014). "Assessing the impact of translation errors on machine translation quality with mixed-effects models." *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, 25-29 October 2014, 1643-1653. <http://www.aclweb.org/anthology/D14-1172> (consulted 18.06.2019).
- **Forcada, Mikel L.** (2017). "Making sense of neural machine translation." *Translation Spaces* 6(2), 291-309.

- **Frawley, William** (1984). "Prolegomenon to a theory of translation." William Frawley (ed.) (1984). *Translation: literary, linguistic and philosophical perspectives*. Newark: University of Delaware Press, 250-263.
- **Froeliger, Nicolas** (2013). *Les Noces de l'analogique et du numérique – De la traduction pragmatique*. Paris: Les Belles lettres, collection "Traductologiques".
- **Guillemin-Flescher, Jacqueline** (1986). *Syntaxe comparée du français et de l'anglais, Problèmes de traduction*. Gap-Paris: Ophrys.
- **Guerberof Arenas, Ana and Joss Moorkens** (2019). "Machine translation and post-editing training as part of a master's programme." *The Journal of Specialised Translation* 31: 217-238.
- **Hassan, Hany et al.** (2018). "Achieving human parity on automatic Chinese to English news translation." <https://www.aclweb.org/anthology/D18-1512.pdf> (consulted 22.10.2019).
- **Hartley, Anthony and Andrei Popescu-Belis** (2004). "Évaluation des systèmes de traduction automatique." Stéphane Chaudiron (ed.) (2004). *Évaluation des systèmes de traitement de l'information*. Paris: Hermès, 311-335.
- **Isabelle, Pierre, Cherry, Colin and George Foster** (2017). "A challenge set approach to evaluating machine translation." *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 27.03.2020)
- **Jia Yanfang, Carl Michael and Xiangling Wang** (2019). "How does the post-editing of neural machine translation compare with from-scratch translation? A product and process study." *The Journal of Specialised Translation* 31: 60-86.
- **Koehn, Philipp** (2010). *Statistical Machine Translation*. Cambridge: Cambridge University Press.
- **Koehn, Philipp and Ulrich Germann** (2014). "The impact of machine translation quality on human post-editing." *Proceedings of the Workshop on Humans and Computer-assisted Translation*, Gothenburg, Sweden, 26 April 2014, 38-46. <http://www.aclweb.org/anthology/W14-0300> (consulted 18.06.2019).
- **Koglin, Arlene and Rossana Cunha** (2019). "Investigating the post-editing effort associated with machine-translated metaphors: a process-driven analysis." *The Journal of Specialised Translation* 31: 38-59.
- **Lapshinova-Koltunski, Ekaterina** (2015). "Variation in translation: evidence from corpora." Claudio Fantinuoli and Federico Zanettin (eds) (2015). *New directions in corpus-based translation studies*. Berlin: Language Science Press, 93-114.
- **Laviosa, Sara** (2002). *Corpus-Based Translation Studies: Theory, Findings, Applications*. Amsterdam/New York: Rodopi.
- **Lin, Chin-Yew** (2004). "ROUGE: a package for automatic evaluation of summaries." *Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004)*, Barcelona, Spain, 25-26 July 2004. <https://www.aclweb.org/anthology/W04-1013.pdf> (consulted 22.10.2019).

- **Loock, Rudy** (2018). "Traduction automatique et usage linguistique : une analyse de traductions anglais-français réunies en corpus." *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." Michel Berré *et al.* (eds) (2019). *La formation grammaticale du traducteur: enjeux didactiques et traductologiques*. Villeneuve d'Ascq: Presses Universitaires du Septentrion, 179-194.
- **Macketanz Vivien et al.** (2017). "Machine translation: Phrase-Based, Rule-Based and Neural approaches with linguistic evaluation." *Cybernetics and Information Technologies* 17(2): 28-43.
- **Martikainen, Hanna et Alexandra Mestivier** (2019). "L'apprenant en traduction face à l'outil nouvelle génération : ses interrogations et espoirs pour l'avenir de la profession traduisante." Paper presented at *Colloque L'apprenant en langues et dans les métiers de la traduction: source d'interrogations et de perspectives* (Université Rennes 2, 31 January-2 February 2019).
- **Massey, Gary and Maureen Ehrensberger-Dow** (2017). "Machine learning: Implications for translator education." *Lebende Sprachen* 62(2): 300-312.
- **Moorkens, Joss** (2018). "What to expect from Neural Machine Translation: a practical in-class translation evaluation exercise." *The Interpreter and Translator Trainer* 12(4): 375-387.
- **Moorkens, Joss et al.** (eds) (2018). *Translation Quality Assessment: From Principles to Practice*. Berlin: Springer.
- **Olohan, Maeve** (2004). *Introducing Corpora in Translation Studies*. London/New York: Routledge.
- **Papineni, Kishore et al.** (2002). "Bleu: a method for automatic evaluation of machine translation." *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 22.10.2019).
- **Popović, Maja et al.** (2013). "Learning from human judgments of machine translation output." *Proceedings of the Machine Translation Summit XIV*, Nice, 2-6 September 2013, 231-238. <http://www.mt-archive.info/10/MTS-2013-Popovic.pdf> (consulted 27.03.2020).
- **Rossi, Caroline** (2017). "Introducing statistical machine translation in translator training: From uses and perceptions to course design and back again." *Revista Tradumàtica. Tecnologies de la Traducció* 15: 48-62.
- **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.
- **Schmid, Helmut** (1994). "Probabilistic part-of-speech tagging using decision trees." *Proceedings of International Conference on New Methods in Language Processing*, Manchester, United Kingdom, 14-16 September 1994. <http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/tree-tagger1.pdf> (consulted 18.06.2019).

- **Screen, Benjamin** (2019). "What effect does post-editing have on the translation product from an end-user's perspective?" *The Journal of Specialised Translation* 31: 133-157.
- **Shterionov, Dimitar et al.** (2018). "Human versus automatic quality evaluation of NMT and PBSMT." *Machine Translation* 32(3): 217-235.
- **Sycz-Opoń, Joanna and Ksenia Gałuskińska** (2017). "Machine translation in the hands of trainee translators – an empirical study." *Studies in Logic, Grammar and Rhetoric* 49(1): 195-212.
- **Toral, Antonio** (2019). "Post-editeese: an exacerbated Translationese." *Proceedings of the Machine Translation Summit XVII*, Dublin, Ireland, 19-23 August 2019, 273-281. <https://arxiv.org/abs/1907.00900> (consulted 11.11.2019).
- **Vanmassenhove, Eva, Shterionov Dimitar and Andy Way** (2019) "Lost in translation: Loss and decay of linguistic richness in machine translation." *Proceedings of the Machine Translation Summit XVII*, Dublin, Ireland, 19-23 August 2019, 222-232. <https://arxiv.org/abs/1907.00900> (consulted 11.11.2019).
- **Vanmassenhove, Eva, Hardmeier Christian and Andy Way** (2018). "Getting gender right in neural machine translation." *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, 31 October-4 November 2018, 3003-3008. <https://www.aclweb.org/anthology/D18-1334> (consulted 18.06.2019).
- **Vinay, Jean-Paul and Jean Darbelnet** (1995). *Comparative Stylistics of French and English: A methodology for translation*. Amsterdam/Philadelphia: John Benjamins.
- **Way, Andy** (2018). "Quality expectations of machine translation: From principles to practice." Joss Moorkens et al. (eds) (2018). *Translation Quality Assessment: From principles to practice*. Cham: Springer, 159-178.
- **Yamada, Masaru** (2019). "The impact of Google Neural Machine Translation on post-editing by student translators." *The Journal of Specialised Translation* 31: 87-106.

Websites

- *2018 European Language Industry Survey Report: Expectations and Concerns of the European Language Industry.* https://euatc.org/wp-content/uploads/2019/10/2018_Language_Industry_Survey_Report.pdf (consulted 18.06.2019).

Biography

Rudy Loock is Professor of English Linguistics and Translation Studies in the Applied Languages Department of the University of Lille, France and affiliated with the CNRS laboratory '*Savoirs, Textes, Langue*'. His research interests include corpus-based Translation Studies, the use of electronic corpora as translation tools, translation quality, as well as translation teaching. He has published a number of articles and book chapters on these topics in English and French, as well as a book entitled *La Traductologie de corpus* (Septentrion, 2016).



Email: rudy.loock@univ-lille.fr

Notes

¹ I would like to sincerely thank the two anonymous reviewers of the article, who have both provided valuable, constructive feedback on a first version, which has led to a better, richer article, with all remaining errors and limitations being naturally my own. Such constructive feedback is not always the norm with all journals unfortunately, so this needs to be explicitly acknowledged.

² The term 'biotranslator' is a derivation from a direct translation of the French neologism 'biotraduction' used for the first time in a 2002 sci-fi novel, *Le Revenant de Fomalhaut* by Jean-Louis Trudel (Froeliger 2013: 20).

³ I would like to thank the European Commission's Directorate-General for Translation for granting me access to eTranslation.

⁴ TSM stands for *Traduction Spécialisée Multilingue*, which is the name of the translation programme at the University of Lille, France, where the (open-ended) corpus is compiled for a comparative grammar class.

⁵ In French, *il y a* can also be used to introduce a period of time, e.g. *il y a 2 ans* ('two years ago'). Such examples are not existential constructions and have not been included in the analysis.