Daems, J., De Clercq, O., & Macken, L. (2017). Translationese and post-editese: How comparable is comparable quality? *Linguistica Antverpiensia, New Series: Themes in Translation Studies*, *16*, 89–103.

Translationese and Post-editese: How comparable is comparable quality?

Joke Daems

Ghent University, Belgium
joke.daems@ugent.be

Orphée De Clercq

Ghent University, Belgium
orphee.declercq@ugent.be

Lieve Macken

Ghent University, Belgium
lieve.macken@ugent.be

Whereas post-edited texts have been shown to be either of comparable quality to human translations or better, one study shows that people still seem to prefer human-translated texts. The idea of texts being inherently different despite being of high quality is not new. Translated texts, for example, are also different from original texts, a phenomenon referred to as 'Translationese'. Research into Translationese has shown that, whereas humans cannot distinguish between translated and original text, computers have been trained to detect Translationese successfully. It remains to be seen whether the same can be done for what we call Post-editese. We first establish whether humans are capable of distinguishing post-edited texts from human translations, and then establish whether it is possible to build a supervised machine-learning model that can distinguish between translated and post-edited text.

1. Introduction

In our increasingly multicultural society, choices need to be made regarding translation production and quality. In order to keep up with the increased need for translation, manual human translation has made way for computer-assisted translation, and – in some circumstances – for the post-editing (PE) of machine-translated texts (Koponen, 2016). Several professional translators are still opposed to the use of machine translation (MT), claiming that it negatively affects the quality of a translation. Research, however, has shown that post-edited (PE) texts are often judged to be of comparable quality to human translations (HT) (Fiederer & O'Brien, 2009; Garcia, 2010; O'Curran, 2014; Plitt & Masselot, 2010) and even of better quality than HTs (Green, 2013; Koponen, 2016). These quality judgements are usually performed by language experts or researchers with a background in linguistics. While they are indeed qualified to perform analyses of textual quality, the perspective of the end-user (the reader) is barely taken into account when judging a text's quality. In fact, to the best of our knowledge, only the research done by Bowker has investigated how recipients of texts evaluate PE and human-translated texts. In 2009, Bowker found that people's tolerance of post-editing and MT depended greatly on the goal of a text and the community under scrutiny, with members of the Fransaskois (a French-speaking Canadian community) greatly preferring HT and West Quebecers mostly preferring PE when they were informed about the production cost and time of HT and PE. A comparable study was performed by Bowker and Buitrago Ciro (2015) with Spanish-speaking immigrants in Canada. They presented readers with different versions of a text (HT, maximally PE, rapidly PE, raw MT) and asked them which text they preferred. Of interest in this study is the fact that the participants first had to give their preference without knowing the source of the text. The respondents chose the HT version of a text in 42% of the cases, compared to

only 24% for the maximally PE texts. This is striking, considering the research into the quality of PE texts. If a fully PE text is indeed of comparable quality to a HT text, what is it that still makes readers prefer HT?

The finding is especially puzzling when compared to the research on Translationese. The term "Translationese" was coined by Gellerstam in 1986, and it has since been used to indicate any type of difference between original text and translated text. In contrast with research on HT and PE texts, user-perception studies are somewhat more common in the field of Translationese. From these studies, it seems that readers are not capable of identifying the difference between an original text and a translated text (Baroni & Bernardini, 2006; Tirkkonen-Condit, 2002). Interestingly, computers have successfully been trained to detect these differences by taking lexical and grammatical information into account (Baroni & Bernardini, 2006; Ilisei, Inkpen, Corpas Pastor, & Mitkov, 2010; Koppel & Ordan, 2011; Volansky, Ordan, & Wintner, 2015).

In this study, we aim to take the first steps towards an identification of what we call "Post-editese": the expected unique characteristics of a PE text that set it apart from a translated text (and, in future work, from original text). The relevance of this work is manifold. Like Translationese, insights into Post-editese can help us to understand both the translation process and the more elusive aspects of translation quality, that is, the aspects of a translated text that make readers prefer it over a PE text of high quality. In the case of Translationese, it seems that despite objective measures of differences between original text and translated text, the intended reader does not usually perceive a difference. In the case of Post-editese, more research is required to investigate further the findings by Bowker (2009) and Bowker and Buitrago Ciro (2015). Some of the more practical applications of Translationese detection as suggested by Baroni and Bernardini (2006) are an assessment tool for translators and translation students, a web-based parallel corpus extractor and multilingual plagiarism detection. A practical application of detecting Post-editese would, for example, be the automatic extraction of non-PE texts to ensure that MT systems are trained on original texts and translations only; another could be a way for post-editors to monitor the output of their work automatically. Considering that PE texts are often of comparable quality to HTs or even of better quality, identifying elements of Post-editese would not necessarily imply identifying elements of lesser quality, but rather identifying those elements that human readers dislike about a PE text that make them prefer an HT text, because this is of importance to people wanting to publish a text.

The research presented in this article attempts to answer two main questions: (1) Can readers spot the difference between HT and PE texts? and (2) Can we identify objective, quantifiable differences between HTs and PE texts? In the following sections, we first elaborate on the importance and features of Translationese and the expected features of Post-editese. This is followed by an outline of the research setup and methodology used, an analysis of our data, and some conclusions and directions for future work.

2. Translationese and Post-editese

While the term "Translationese" has been used to denote bad translation, Gellerstam (1986) originally intended it to mean statistical differences between translated and original text. Baker (1993) introduced the notion of translation universals: typical features of translation, independent of language combination. She proposed four such translation universals: simplification, explicitation, normalization and interference. Simplification means that complex features are replaced by simpler features in a translated text; explicitation means that implicit information is made explicit more often in a translated text; normalization means that translated texts are often more standardized, using conventional grammar; and interference means that the source language's (SL) influence is visible in the translation. Corpus studies tried to find proof of these universals by, for example, looking at the type–token ratio (lexical variety) (Al-Shabab, 1996), sentence length and the ratio of content to non-content words (lexical density) (Laviosa, 1998) in translated text.

More recently, machine-learning strategies have been used to identify differences between translated and original texts, which has also led to the notion of translation universals being challenged. Volansky et al. (2015), for example, established that some of the characteristics of

translation depend greatly on the language pair. Baroni and Bernardini (2006) were, to the best of our knowledge, the first to use support vector machines (SVMs) to identify translated texts. They found that function words, personal pronouns and adverbs are some of the main features used by the SVMs to identify translated Italian. Ilisei et al. (2010) found proof for the simplification universal in Spanish, also using SVMs. Their system relied heavily on lexical richness, the proportion of grammatical words to lexical words, sentence length, word length and – compared to what Baroni and Bernardini (2006) found – morphological attributes. The previous two studies were examples of supervised machine-learning studies. Rabinovich and Wintner (2015) successfully applied unsupervised machine learning to the identification of Translationese, mostly using function words, character trigrams and part-of-speech (PoS) trigrams.

As this is, to the best of our knowledge, the first article to consider the possible features and perceptions of what we will call "Post-editese", our assumptions are naturally limited to what we know about Translationese and PE in general. Where we expect there to be source text (ST) interference in Translationese, we expect there to be MT interference in Post-editese, as post-editors are primed by the MT output (Green et al., 2013). Aharoni, Koppel and Goldberg (2014) were able to automatically identify sentences as being MTs or HTs, using features such as PoS and information about function word frequency. Lapshinova-Koltunski (2013) built a corpus containing HT texts, various types of MT and computer-assisted translation. She managed to discriminate between HTs and MT on the basis of conjunctions, personal pronouns and adverbs. Verbs, adjectives and nouns helped to discriminate between three groups: computer-assisted translation and rule-based MT, HT and statistical MT. There therefore seems to be a type of Machine Translationese, although the question remains whether its features can also be found in Post-editese. The only study moving in the direction of identifying Post-editese is that by Čulo and Nitzke (2016): they compared the terminology used in MT, PE texts and HT and found that the PE terminology was closer to that of the MT output than to that of the HT.

3. Corpus collection and processing

The research presented in this article comprises two studies: a reader-perception study in which participants had to label texts as being either PE or HT, and a quantitative study in which textual information was analysed across translation methods. The main goal was to identify whether translations and PE texts of publishable quality still exhibit (perceived) unique characteristics that set them apart from one another.

The corpus was collected during a previous study (Daems, 2016), in which 13 professional translators (age range 25–51) and 10 master's students of translation (age range 21–25) post-edited and translated eight different newspaper articles of approximately 150–160 words long from English into Dutch. The goal in both tasks was to obtain a text of publishable quality. With the exception of one translator, who had two years of experience, all the translators had a minimum of five years and a maximum of 18 years of experience working as a full-time professional translator. The students had all passed their final English Translation examination. The participants had limited to no experience with PE. Text topics varied for each text: for example, from "the impact of climate change on violence" to "criticism on using lie detector tests in job application procedures". For a full discussion of how the texts were selected as well as an overview of the different texts, see Daems (2016). After discarding incomplete data, the corpus consisted of 87 translations and 87 PE Dutch texts (10 to 11 versions of each source text, approximately half of which were made by each participant group). The study was approved by the Ethical Commission of the Faculty of Psychology and Educational Sciences at Ghent University. All the participants gave their written informed consent.

The translations and PE texts in the original study were manually annotated by two of the authors of this article using a two-step translation quality-assessment approach[1] (Daems, Macken, & Vandepitte, 2013). This approach takes two aspects of quality into account: acceptability, or adherence to target norms, language, and structure, on the one hand, and adequacy, or a comparison of ST and target text (TT), on the other, to see whether the information contained in the first was still

present and unchanged in the latter. The annotators first annotated the text for acceptability by looking at the TT only, then annotated the text for adequacy by considering both the ST and the TT in parallel. After annotation, a consolidation phase took place, during which the annotators discussed the annotations they did not agree on. Inter-annotator agreement was calculated during pretests of the method, showing a high level of agreement between annotators after consolidation (from 67% with $\kappa = .65$ in an earlier experiment to 95% with $\kappa = .94$ in a later pretest). Only the annotations that both annotators agreed on after consolidation have been used for further analysis. Both the acceptability and the adequacy categories contain a variety of subcategories that receive error weights depending on the severity of the error (for example, the acceptability subcategory "capitalization error" receives an error weight of 1, whereas the adequacy subcategory "contradiction" receives an error weight of 4). The average error weight (EW) per word was calculated for each translation and PE text. A linear mixed effects model[2] with average error weight as dependent variable and translation method (HT and PE) as predictor variable did not outperform the null model, indicating that there is no statistically significant difference in quality between the HTs and the PE texts in the corpus.

After creating the corpus, we selected the texts to be used in both studies. In order to have as many data points as possible, the whole corpus was used to perform the quantitative study. For the reader perception study, a subset of the corpus was used in order to have multiple reader evaluations for each text. To create the subset, we selected the two translated versions and two PE versions with the highest quality for each of the eight source texts, regardless of the participant group. Highest quality was determined by the lowest average EW per word.

Table 1 shows information on the average EW, across all the texts and across the selected texts only. As can be seen, the average EWs of the selected texts are well below those of the full text set. To verify that the high quality of the PE texts was not simply due to the translators' deleting the MT output and creating their own translation from scratch, we calculated the Translation Edit Rate (TER) on the PE texts. TER measures the edit distance between the MT output and the final PE text, using a score from 0 to 100, with a lower TER score meaning that fewer edits are needed to turn an MT sentence into the final PE sentence. While TER is not an indication of the actual editing effort, it is an indication of the correspondence between the MT output and the final PE product, regardless of how the translation was produced. As we were looking for Post-editese in a finished text only, and we expected Post-editese to manifest itself through priming from the MT output, the most important parameter is the amount of overlap between MT output and the PE product. As such, it does not matter whether that priming was caused by post-editing only select parts of the MT output or by typing a new translation that was heavily primed by the MT output. Both are expected to exhibit comparable characteristics of Post-editese. As can be seen in Table 2, the edit rate of the selected texts is comparable to that of the rest of the texts, and is never higher than 74.3%. Figure 1 shows the distribution of TER values across all PE texts.

Table 1: Comparison of average EW for all texts and for the subset used for the reader perception study.

	EW min	EW max	EW mean	EW median
All texts	0	0.167	0.051	0.048
Subset	0	0.066	0.015	0.011

Table 2: Comparison of the TER for all PE texts and for the subset used for the reader perception study.

	TER min	TER max	TER mean	TER median
All PE texts	26.9	76.3	52.3	52.1
Subset	40.6	74.3	58.4	60.7

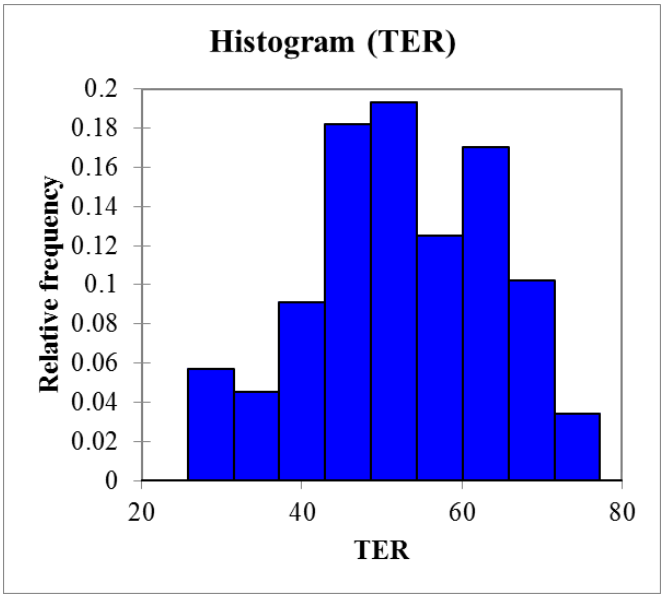


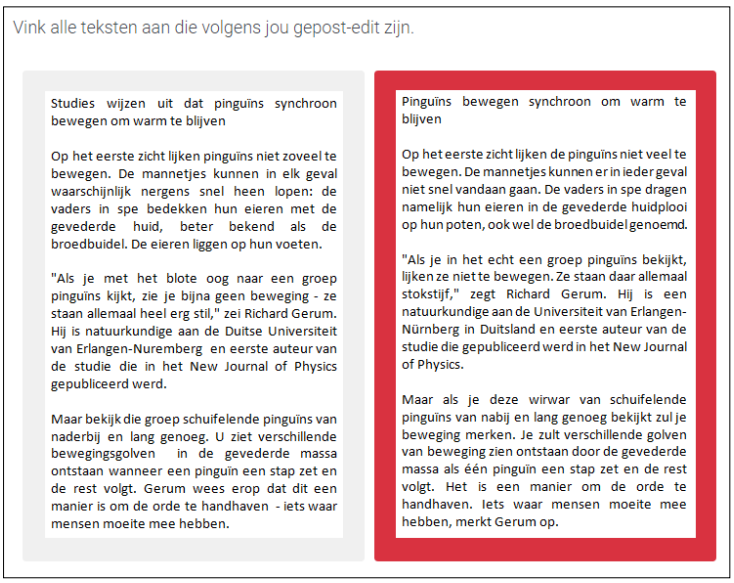
Figure 1: Distribution of TER values across PE texts.

4. Reader perception study

4.1 Survey

A survey was created using the Qualtrics online data-collection software (Qualtrics, Provo, UT). We converted the 32 texts (two HT versions and two PE versions for each of the eight source texts) to images in order to be able to integrate them in a graphic horizontal multiple-choice question and to ensure that the formatting would stay consistent across devices. Each question showed the participant two text versions of the same source text in parallel. An example question is shown in Figure 2.

Figure 2: Example of graphic horizontal multiple-choice question.



The question was always 'mark the texts you think are PE'. The participants could choose to select one text, two texts or no texts. The main question was followed by a question for additional information, where the participants had to explain why they had made the choice they had. In order to prevent influence from seeing the same text more than once and to counter possible fatigue effects, each participant was presented with four different questions only (from four different source texts). There were six different text combinations for each source text: two HT texts, two PE texts and four ways in which a PE text could be presented together with an HT text (PE1HT1, PE2HT1, PE1HT2, PE2HT2). The survey setup consisted of eight blocks, one for each source text. In order to counter task-order effects and to collect a comparable amount of data across all texts and conditions, block randomization was added to Qualtrics, with a selection of four blocks, that is, source texts, per participant, and question randomization, with one question randomly selected from the six possible text combinations. The position of the text images on the screen (either left or right) was also randomized automatically by Qualtrics.

4.2 Participants

The survey was presented to two groups of translation students at Ghent University as part of their courses on Introduction to Translation Technology, Terminology and Translation Technology, and Machine Translation and Post-editing, and was shared with people working at the Translation department via email. A total of 195 people completed the survey. Ages ranged from 18 to 64, with most participants (135) falling in the 18–22 range.

4.3 Data analysis

Data was collected from 18 October to 3 November 2016. Of the 195 surveys received, 174 were filled in completely and were therefore retained for the analysis.

The main goal of the survey was to answer the question: "Are people capable of identifying a text as being PE or being translated from scratch?" We looked at the data in two ways: per text

combination, and per text. For the first analysis, we looked at the four possible ways in which texts could be presented (HT-HT, PE-PE, PE-HT, HT-PE) and the corresponding labels participants assigned to the two texts (HT-HT, PE-PE, PE-HT, HT-PE). We then checked how often the correct condition was assigned to each set.

For the second analysis, we looked at individual text assessments. A text could either be HT or PE, and we checked whether the label assigned by the participants (HT or PE) corresponded to the actual text-production method. The results are presented in contingency tables. To assess the results statistically, we calculated precision and recall for the different tables.

4.4 Results

Tables 3 and 5 are contingency tables that show the actual labels of the conditions and texts alongside the labels assigned by the participants. As can be derived from Table 3, the participants assigned the correct labels in just less than 30% of the cases ((13 + 16 + 90 + 87)/694 × 100). This means that, in contrast to the findings by Bowker and Buitrago Ciro (2015), and more in line with the research on Translationese (Baroni & Bernardini, 2006), readers do not seem to experience a difference between HTs and PE texts.

Table 3: Contingency table per text set. (Correctly assigned labels are marked in italics.)

		Actual text displayed			
		PE-PE	HT-HT	PE-HT	HT-PE
Assigned by participants	PE-PE	*13*	11	21	23
	HT-HT	16	*16*	31	36
	PE-HT	49	45	*90*	89
	HT-PE	38	42	87	*87*

Interestingly, PE texts in the PE-PE condition *and* the PE-HT condition are more often incorrectly labelled as being HTs than HT texts are incorrectly labelled as being PE. These findings are reflected in the precision and recall scores, summarised in Table 4. It is striking that the PE-PE (13, 11, 21, 23) and HT-HT (16, 16, 31, 36) conditions are chosen much less frequently than the PE-HT (49, 45, 90, 89) and HT-PE (38, 42, 87, 87) conditions (Table 3) and that they also had worse results overall (Table 4).

Table 4: Overview of precision and recall for each text set condition.

Text set condition	Precision	Recall
PE-PE	19.118%	11.207%
HT-HT	16.162%	14.035%
PE-HT	32.967%	39.301%
HT-PE	34.252%	37.021%

In Table 5, we see that, for the individual text labels, correct and incorrect labels are almost equally common for HT and PE texts. Again, there seems to be a tendency for the participants to select HT more often than PE.

Table 5: Contingency table per individual text. (Correctly assigned labels are marked in italics.)

		Actual conditions	
		HT	PE
Assigned by participants	HT	*363*	364
	PE	331	*334*

Table 6: Overview of precision and recall for individual text labels.

Text label	Precision	Recall
HT	49.931%	49.931%
PE	50.226%	47.851%

The high level of incorrect labels is also reflected in low precision and recall here (see Table 6). This again seems to indicate that the participants are not capable of correctly distinguishing between HTs and PE texts.

5. Computational analysis

Whereas the first study showed that humans are not capable of distinguishing between both types of text, we were also interested in verifying whether a computer can identify the difference. Various studies have shown that it is possible to identify Translationese (differences between original text and translated text) using supervised machine-learning techniques (Baroni & Bernardini, 2006; Ilisei et al., 2010; Koppel & Ordan, 2011; Volansky et al., 2015). In this section, similar experiments are performed. A first prerequisite is to linguistically process all 174 texts in our corpus and derive text characteristics or features. For this feature extraction we were inspired by the readability prediction system developed by De Clercq and Hoste (2016) and previous work on Translationese.

5.1 Feature extraction

We implemented different types of text characteristic, amounting to 55 distinct features. The features can be divided in four groups: traditional,[3] lexical, syntactic and semantic. All of these features were computed at the text level using state-of-the-art text-processing tools, as explained below. The decision was made to include these four feature groups based on previous research on Translationese, the intuition being that traditional and lexical features are related to the translation universal of simplification, syntactic features can give an indication of interference, and semantic features, in particular cohesive markers, are relevant to identifying explicitation.

The traditional features include four length-related features that have proved successful in readability prediction research (François & Miltsakaki, 2012): average word and sentence length, ratio of long words in a text (i.e. words containing more than three syllables) and percentage of polysyllabic words. These features were obtained after processing the texts with the Dutch preprocessor Frog (Van den Bosch et al., 2007) and a designated classification-based syllabifier (Van Oosten, Tanghe, & Hoste, 2010). Next, a number of lexical features were calculated, including the percentage of words that can be found in the CLIB list (Staphorsius, 1994), which comprises the most frequently used words in Dutch, and the type–token ratio in order to measure the lexical complexity within a text. Besides these easy-to-calculate features, we also incorporated more advanced features inspired by work on language modelling and terminology extraction. Both feature types are based on a reference corpus, in our case the SoNaR corpus (Oostdijk, Reynaert, Hoste, &

Schuurman, 2013). Because we were working with edited text, we derived a subset of this large reference corpus that comprises only text from edited genres: newspaper, magazine and Wikipedia material. Two language-modelling features were included: one where the perplexity of a given text when compared to a reference corpus is given (perplex) and another where this score was normalized over the text length (normperplex). The Term Frequency-Inverse Document Frequency, tf-idf (Salton, 1989) and the Log Likelihood (Rayson & Garside, 2000) ratio of all the terms included in a particular text were included as terminological features.

Next, we incorporated two types of syntactic features: a shallow level, where all the features are computed based on parts of speech (PoS) tags, and a deeper level based on dependency parsing. Based on the PoS, we first incorporated two overall features: the average number of content and function words within a text. Next, 25 features were calculated based on the following five PoS: nouns, adjectives, verbs, adverbs and prepositions. We indicated the absolute and relative frequency for each class in the text and in the sentence, as well as the average type per sentence as determined using the Frog preprocessor. For the next phase, however, we used the Alpino dependency parser for Dutch (Van Noord et al., 2013) to parse all the texts and calculated the average parse tree height, number of subordinating conjunctions, number of passive constructions and the ratio of the noun, verb and prepositional phrases.

Lastly, we also incorporated some basic semantic features based on lists of connectives since these serve as an important indication of text cohesion in a text (Halliday & Hasan, 1976). These lists were drawn up by a linguistics expert (Denturck, 2014). As features, we counted the average number of connectives within a text and the average number of causal, temporal, additive, contrastive and concessive connectives at both the sentence and the text level.

All the features were used in the experiments.

5.2 Experimental Design

As mentioned in Section 1, all available texts were used for the experiments. This means we have a dataset of 174 texts available for our experiments with an equal class distribution: 87 PE texts and 87 HTs. In order to perform supervised machine-learning experiments this dataset was subdivided into a 90% train and a 10% test split, following the same class distribution. This resulted in 158 texts for training and 16 texts for testing. The selection of test texts was also influenced by the decision to include an equal number of high-quality and low-quality texts based on the average EW per word (see Section 1), since this might have offered insight into our models.

Our main research question is: Is it possible to build a supervised machine-learning model that can distinguish between translated and PE text? For the research presented here, this boils down to a binary classification task: PE (label “1”) or translated (label “0”). We are equally interested in discovering whether features modelling lexical, syntactic and semantic text characteristics are up to the task and, if so, which features contribute most. To this purpose, we performed two different rounds of experiments.

In Round 1, we first examined the individual feature contributions in our training data. It is possible to compute statistics about the relevance of features by looking at those features that are good predictors of the class labels based on Information Theory (Quinlan, 1986). Information Gain (IG) weighting looks at each feature in isolation and measures how much information it contributes to our knowledge of the correct class label. This statistic, however, tends to overestimate the relevance of features with large numbers of values, which is why IG is often reported together with Gain Ratio (GR), its normalized version (Quinlan, 1993). In subsequent work, White and Liu (1994) have shown that the GR measure still has an unwanted bias towards features with more values, and propose the chi-squared statistic as an alternative. We calculated all three statistics on our training dataset. The resulting values can be interpreted as feature weights and ranked according to the amount of information they add to discriminating between the two possible labels. Next, we also tried to fit a logistic regression model to our training data in order to discover which features contribute most. Finally, this model was also tested on our held-out test set.

In these first experiments, all the features were considered independently of one another. This is not necessarily the best strategy and often better results can be obtained by leaving features out and focusing more on the feature interplay. That is why, in Round 2, we switched to a more advanced technique by exploiting a wrapper-based approach to feature selection using genetic algorithms. In a wrapper approach, feature informativeness is determined while running an induction algorithm on a training dataset and the best features are selected in relation to the problem to be solved. Finding a good subset of features requires searching the space of feature subsets. We used genetic algorithms (GAs) for this purpose and ran tenfold cross-validation on the training data (see Mitchell, 1996 for more information on genetic algorithms). We used TiMBL (Daelemans, Zavrel, Van der Sloot, & Van den Bosch, 2010) as our classifier, a nearest neighbour algorithm, ensuring that $k = 1$ because we were dealing with a small dataset. To evaluate, we calculated accuracy. For the optimization experiments, we allowed for individual feature selection, which should enable us to visualize those features, and especially those feature interplays, that contributed most to the classification tasks. We started from a population of 100 individuals and allowed 100 generations. We set the stopping criterion to a best fitness score (accuracy) that remained the same during the last five generations. All the optimization experiments were performed using the Gallop toolbox (Desmet, Hoste, Verstraeten, & Verhasselt, 2013), which is specifically aimed at natural language.

5.3 Results Round 1

Based on our training data, we calculated IG, GR and chi-squared. These values can be interpreted as feature weights and ranked according to the amount of information they add to discriminating between the two possible labels: PE versus HT. Table 7 presents the top ten features according to all three statistics.

Table 7: Top ten features according to three statistics from Information Theory: information gain (IG), gain ratio (GR) and chi-squared (× 2).

IG	GR	X2
Avg word length	Avg word length	Avg word length
Avg tfidf	Avg tfidf	Avg tfidf
Avg LL	Avg LL	Avg LL
Perplexity	Perplexity	Perplexity
Normalized perplexity	Normalized perplexity	Normalized perplexity
Ratio long words	Ratio of long words	Ratio long words
Type-token Ratio	Type-token Ratio	Type-token Ratio
% frequent DU	% frequent DU	% frequent DU
% polysyllable words	Avg noun types	% polysyllable words
Avg nouns	Avg nouns	Avg nouns

From the results we observe that all three statistics more or less agree on which features are most discriminative; these are indicated in italics. These comprise all of the lexical features (percentage of frequent Dutch words, type-token ratio, average tf-idf and log-likelihood score and both language modelling features), two traditional features related to length (average word length, ratio of long words) and one shallow syntactic feature (average number of nouns).

These statistics, however, do not give much insight into whether a model would actually be able to discern PE from translated text. To investigate this we attempted to fit a logistic regression model onto our training data. Inspection of the model fit provides a closer look at those coefficients (features) that are considered statistically significant variables. We also analysed the table of deviance in a subsequent phase. The features that were found to be statistically significant are presented in Table 8.

Table 8: Statistically significant variables according to the logistic regression model (left) and the table of deviance (right). (Features common to both lists are indicated in italics. Asterisks denote significance of results: * = $p \leq 0.05$; ** = $p \leq 0.01$; *** = $p \leq 0.001$)

Coefficients		Deviance	
Feature	p-value	Feature	p-value
avg_adj_types	0.000250***	np_count	0.007383**
avg_verb_types	0.001429**	avgtfidf	0.007767**
Type-token ratio	0.003396**	avg_adj_types	0.009645**
avg_type_adj	0.004996**	perplex	0.011645*
Avg_type_verb	0.005233**	avg_prep_sent	0.02794*
Avg_adverb_types	0.044568*	parse_tree_depth	0.037456*
		avg_verb_types	0.038163*

When comparing the two parts of the table, we see that different features are indicated. The only features that occur in both lists are the average number of adjective types and the average number of verb types (both indicated in italics). These are both shallow syntactic features based on PoS tagging information. Actually, if we consider the coefficients only, all but one are derived from PoS information. The deviance scores, on the other hand, tell a different story. The feature allowing for the highest residual deviance in comparison to the null model is the average number of noun phrases, a complex syntactic feature, followed closely by the average tf-idf value.

Next, we tested our fitted model on our held-out test set to see whether our model was actually able to generalize to unseen data. This resulted in an accuracy of 56.23%. Comparing this to a baseline relying only on the even class distribution (50%), we can conclude that our model has actually learnt something.

Based on these analyses and the performance gain over the baseline when testing the model on our reserved test set, we could conclude that a classifier can learn to distinguish between PE and HT text when assigning most weight to lexical and syntactic features. However, the performance gain over the baseline is very moderate and for these experiments all the features were still included in the model, which is not necessarily the best choice. This brings us to the second round of experiments.

5.4 Results Round 2

In Table 9 we compare our baseline with ten-fold cross validation experiments on the training data. In the first setting we simply used all available features, whereas in the second setting we performed the optimization experiments as explained in Section 5.2.

Table 9: Results of the tenfold cross-validation experiment on the training data, represented by accuracy.

Setting	Accuracy
Baseline	50.00
All features	51.26
Optimization	68.31

These results are promising, especially those from the optimization experiments, where accuracy improves by no less than 18 points. An interesting part of the Gallop toolkit is that it also offers its users insight into those features that either were or were not selected in the fittest individuals. For the

present experiment, 31 of the 55 features were selected. Of the traditional features, three were selected (average word length, ratio of long words and percentage of polysyllabic words). Examining the lexical features, the two language-modelling features (perplexity and normalized perplexity) were selected, as was the average tf-idf value. As for the syntactic features, the two more global features representing the average number of content and function words were retained, as well as one feature relating to the PoS category noun (average type nouns), four features relating to the adjectives, and three features each relating to verbs, adverbs and prepositions. Regarding the more complex syntactic features, based on dependency parsing, the numbers of noun phrases, verb phrases and passives were also considered important. Finally, regarding the shallow semantic features, the average number of connectives at the sentence level is maintained, as are those features that indicate causal, additive, contested or concessive relations.

This leads us to conclude that for this particular task all of the different feature types seem to contribute to the actual performance. However, a problem that often occurs when performing cross-validation experiments on training data is that of overfitting. Therefore, it is important also to test the final model on a held-out test set. When we tested our model using all the available features, which achieved an accuracy of 51.26 on our training data, the accuracy level dropped to 50.00% when testing on the held-out test set; this is the same as our baseline. When we did the same with our optimal model and trained and tested only including the selected features, the performance dropped dramatically from 68.31 to 43.75 on our held-out test set. This leads us to conclude that it is not possible to create a classifier that is able to distinguish between PE and translated text in the current setup. Whether this is due to the feature representations or the low amount of training data is something that will have to be explored in future research.

Table 10: Features that were and were not selected in the optimal setting on training data.

average_word_length	1	avg_adj_sent	1	avg_conn_doc	0
average_sentence_length	0	avg_type_adj_sent	1	avg_conn_sent	1
ratio_long_words	1	avg_adj_types	0	avg_cause_doc	1
percentage_polysyllable_words	1	avg_verb	1	avg_cause_sent	0
percentage_frequent_nl_words	0	avg_type_verb	1	avg_temp_doc	0
type_token_ratio	0	avg_verb_sent	0	avg_temp_sent	0
Avgtfidf	1	avg_type_verb_sent	1	avg_add_doc	0
Avgll	0	avg_verb_types	0	avg_add_sent	1
Perplex	1	avg_adverb	0	avg_cont_doc	1
Normperplex	1	avg_type_adverb	1	avg_cont_sent	1
avg_content	1	avg_adverb_sent	0	avg_conc_doc	1
avg_funct	1	avg_type_adverb_sent	1	avg_conc_sent	0
avg_noun	0	avg_adverb_types	1	parse_tree_depth	0
avg_type_noun	1	avg_prep	1	sbar_count	0
avg_noun_sent	0	avg_type_prep	0	np_count	1
avg_type_noun_sent	0	avg_prep_sent	0	vp_count	1
avg_noun_types	0	avg_type_prep_sent	1	pp_count	0
avg_adj	1	avg_prep_types	1	passives	1
avg_type_adj	1				

6. Conclusion

We did not find proof of the existence of Post-editese, either perceived or measurable.

The user perception study showed that the participants were unable to distinguish between HT and PE texts of publishable quality. If anything, they more often incorrectly labelled PE texts as HTs than the other way around. This is in contrast to the findings by Bowker and Buitrago Ciro (2015) that readers had a clear preference for HT, even when they did not know how a translation was produced. As indicated by the Bowker (2009) study, different language communities have different attitudes towards MT and PE, and it is possible that our findings can be attributed to the different language combination (English–Dutch). Our findings are also more in line with those from Translationese research, where readers were unable to distinguish between translated and original texts (Baroni & Bernardini, 2006; Tirkkonen-Condit, 2002). It was striking that participants more often thought that the two presented texts were from different conditions (HT-PE or PE-HT) rather than from the same condition (HT-HT or PE-PE). Perhaps this was caused by the fact that two texts were presented on screen and the participants involuntarily felt that they had to find differences between the two texts.

The computational analysis seemed promising at first, with a variety of features and combinations of features seemingly being able to help discriminate between HT and PE. Some of the promising features correspond to features also found to be useful in related work: sentence length (Ilisei et al., 2010), perplexity (Čulo & Nitzke, 2016), average amount of content and function words (Ilisei et al., 2010; Laviosa, 1998; Rabinovich & Wintner, 2015), and conjunctions (Lapshinova-Koltunski, 2013), among others. After testing the suggested models on a held-out dataset, however, performance showed that, like humans, the computer is not capable of accurately distinguishing between HT and PE.

Our findings could be an indication that there is indeed no such thing as "Post-editese" and that fully PE texts are indistinguishable from HT texts with regard to quality, reader perception, and traditional, lexical, syntactic and semantic features. Different results can be expected for texts of varying levels of quality, but this study was concerned with identifying possible Post-editese in a high-quality scenario to see whether a reader would be able to identify a publishable text as being PE or not, so that the comparison with the Bowker and Buitrago Ciro (2015) study could be made. While there was no measurable difference in quality between the texts produced by professional translators and students, there could be other differences between both, and those differences may have had an impact on the identification of Post-editese. Alternatively, our findings could be due to the text type and language combination. The computational results in particular have to be interpreted with caution. Though the genetic algorithm is computationally highly advanced, the current dataset is rather small. The lack of significant results on the held-out data could simply be a consequence of insufficient training data in general.

In future work, our analyses should be repeated on a larger dataset and tested on a variety of text genres and language combinations. Depending on the goal of the evaluation, texts of lower quality could be compared to see whether Post-editese is more evident for lower-quality texts. The user perception study could be improved by either only presenting one text on screen at a time or by introducing control trials with two texts that are exactly the same to ensure that the participants are engaged in the task. An additional factor to control for in future work would be the post-editor, by looking at experience or PE strategies in addition to the level of quality we had already controlled for.

References

Aharoni, R., Koppel, M., & Goldberg, Y. (2014, June). *Automatic detection of machine translated text and translation quality estimation.* Paper presented at the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers), Baltimore, MD.

Al-Shabab, O. (1996). *Interpretation and the language of translation: Creativity and conventions in translation*. Edinburgh: Janus.

Baker, M. (1993). Corpus linguistics and translation studies: Implications and applications. In M. Baker, G. Francis, & E. Tognini-Bonelli (Eds.), *Text and technology: In honour of John Sinclair* (pp. 233–252). Amsterdam: John Benjamins.

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

Bowker, L. (2009). Can Machine Translation meet the needs of official language minority communities in Canada?: A recipient evaluation. *Linguistica Antverpiensia New Series – Themes in Translation Studies, 8*, 123–155.

Bowker, L., & Buitrago Ciro, J. (2015). Investigating the usefulness of machine translation for newcomers at the public library. *Translation and Interpreting Studies, 10*(2), 165–186.

Čulo, O., & Nitzke, J. (2016). Patterns of terminological variation in post-editing and of cognate use in machine translation in contrast to human translation. *Baltic Journal of Modern Computing, 4*(2), 106–114.

Daelemans, W., Zavrel, J., Van der Sloot, K., & Van den Bosch, A. (2010). *TiMBL: Tilburg Memory Based Learner, version 6.3, Reference Guide*.

Daems, J. (2016). *A translation robot for each translator?: A comparative study of manual translation and post-editing of machine translations: process, quality and translator attitude.* Ghent University. Faculty of Arts and Philosophy, Ghent, Belgium.

Daems, J., Macken, L., & Vandepitte, S. (2013). Quality as the sum of its parts: A two-step approach for the identification of translation problems and translation quality assessment for HT and MT+PE. *Proceedings of MT Summit XIV Workshop on Post-editing Technology and Practice, Nice, France*, *63–71*.

De Clercq, O., & Hoste, V. (2016). All mixed up?: Finding the optimal feature set for general readability prediction and its application to English and Dutch. *Computational Linguistics, 42*(3), 457–490.

Denturck, K. (2014). *Et pour cause...: la traduction de connecteurs causaux à la lumière des universaux de traduction: Une étude de corpus (français–néerlandais, néerlandais–français)*. Ghent University. Faculty of Arts and Philosophy, Ghent, Belgium.

Desmet, B., Hoste, V., Verstraeten, D., & Verhasselt, J. (2013). *Gallop Documentation*. Retrieved from https://www.lt3.ugent.be/publications/gallop-documentation/

Fiederer, R., & O'Brien, S. (2009). Quality and Machine Translation: A realistic objective? *The Journal of Specialised Translation, 11*, 52–74.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, 32(3), 221–233.

François, T., & Miltsakaki, E. (2012, June). *Do NLP and machine learning improve traditional readability formulas?* Paper presented at the 1st Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR2012), Montreal, QC.

Garcia, I. (2010). Is machine translation ready yet? *Target, 22*(1), 7–21.

Gellerstam, M. (1986, June). *Translationese in Swedish novels translated from English*. Paper presented at the Scandinavian Symposium on Translation Theory, Lund.

Green, S., Heer, J., & Manning, C. (2013, May). *The efficacy of human post-editing for language translation*. Paper presented at the ACM Human Factors in Computing Systems (CHI), Paris.

Halliday, M., & Hasan, R. (1976). *Cohesion in English*. Longman Group.

Ilisei, I., Inkpen, D., Corpas Pastor, G., & Mitkov, R. (2010). Identification of Translationese: A machine learning approach. In A. Gelbukh (Ed.), *Computational linguistics and intelligent text processing: 11th International Conference, CICLing 2010, Iaşi, Romania, 21–27 March 2010. Proceedings* (pp. 503–511). Berlin: Springer.

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

Koppel, M., & Ordan, N. (2011, June). *Translationese and its dialects.* Paper presented at the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, OR.

Lapshinova-Koltunski, E. (2013, August). *VARTRA: A comparable corpus for analysis of translation variation.* Paper presented at the 6th Workshop on Building and Using Comparable Corpora, Sofia, Bulgaria.

Laviosa, S. (1998). Core patterns of lexical use in a comparable corpus of English lexical prose. *Meta, 43*(4), 557–570.

Mitchell, M. (1996). *An introduction to genetic algorithms*. MIT Press, Cambridge.

O'Curran, E. (2014, October). *Translation quality in post-edited versus human-translated segments: A case study*. Paper presented at the AMTA 2014 3rd Workshop on Post-editing Technology and Practice (WPTP-3), Vancouver, BC.

Oostdijk, N., Reynaert, M., Hoste, V., & Schuurman, I. (2013). The construction of a 500-million-word reference corpus of contemporary written Dutch. In P. Spyns & J. Odijk (Eds.), *Essential speech and language technology for Dutch: Theory and applications of natural language processing* (pp. 219–247). Springer, Berlin.

Plitt, M., & Masselot, F. (2010). A productivity test of statistical machine translation: Post-editing in a typical localisation context. *The Prague Bulletin of Mathematical Linguistics, 93*,7–16.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning, 1*, 81–106.

Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. San Mateo, CA: Morgan Kaufmann.

Rabinovich, E., & Wintner, S. (2015). Unsupervised identification of translationese. *Transactions of the Association for Computational Linguistics, 3*, 419–432.

Rayson, P., & Garside, R. (2000, October). *Comparing corpora using frequency profiling.* Paper presented at the 38th Annual Meeting of the Association for Computational Linguistics Workshop on Comparing Corpora, Hong Kong.

Salton, G. (1989). *Automatic text processing: The transformation, analysis and retrieval of information by computer.* Addison-Wesley Longman, Reading.

Staphorsius, G. (1994). *Leesbaarheid en leesvaardigheid: De ontwikkeling van een domeingericht meetinstrument*. Universiteit Twente.

Tirkkonen-Condit, S. (2002). Translationese - a myth or an empirical fact?: A study into the linguistic identifiability of translated language. *Target, 14*(2), 207–220.

van den Bosch, A., Busser, B., Daelemans, W., & Canisius, S. (2007, December). *An efficient memory-based morphosyntactic tagger and parser for Dutch.* Paper presented at the Seventeenth Computational Linguistics in the Netherlands (CLIN), Nijmegen.

van Noord, G. J. M., Bouma, G., van Eynde, F., de Kok, D., van der Linde, J., Schuurman, I., Sang, E. T. K., & Vandeghinste, V. (2013). Large scale syntactic annotation of written Dutch: LASSY. In P. Spyns & J. Odijk (Eds.), *Essential speech and language technology for Dutch: Theory and applications of natural language processing* (pp. 231–254). Heidelberg: Springer.

van Oosten, P., Tanghe, D., & Hoste, V. (2010, May). *Towards an improved methodology for automated readability prediction.* Paper presented at the 7th International Conference on Language Resources and Evaluation (LREC-2010), Valletta.

Volansky, V., Ordan, N., & Wintner, S. (2015). On the features of translationese. *Digital Scholarship in the Humanities, 30*(1), 98–118.

White, A. P., & Liu, W. Z. (1994). Bias in information-based measures in decision tree induction. *Machine Learning, 15*(3), 321–329.

---

1 http://users.ugent.be/~jvdaems/TQA_guidelines_2.0.html

2 The average EW was right-skewed because many of the sentences contained no errors, leading to a high number of zero values. No transformation was performed, because these values form an integral part of the data and could not be meaningfully interpreted otherwise. Fixed effects in linear mixed models are, moreover, robust to deviations from the normality assumption.

3 The term “traditional” is chosen to refer to those text characteristics that have been used in the first systems to measure the readability of a text, namely readability formulas, such as the well-known Flesch Reading Ease (Flesh, 1948).