Translationese Features as Indicators of Quality in English-Russian Human Translation

Maria Kunilovskaya
University of Tyumen
University of Wolverhampton
maria.kunilovskaya@wlv.ac.uk

Ekaterina Lapshinova-Koltunski
Saarland University
e.lapshinova@mx.uni-saarland.de

Abstract

We use a range of morpho-syntactic features inspired by research in register studies (e.g. Biber, 1995; Neumann, 2013) and translation studies (e.g. Ilisei et al., 2010; Zanettin, 2013; Kunilovskaya and Kutuzov, 2018) to reveal the association between translationese and human translation quality. Translationese is understood as any statistical deviations of translations from non-translations (Baker, 1993) and is assumed to affect the fluency of translations, rendering them foreign-sounding and clumsy of wording and structure. This connection is often posited or implied in the studies of translationese or translational varieties (De Sutter et al., 2017), but is rarely directly tested. Our 45 features include frequencies of selected morphological forms and categories, some types of syntactic structures and relations, as well as several overall text measures extracted from Universal Dependencies annotation. The research corpora include English-to-Russian professional and student translations of informational or argumentative newspaper texts and a comparable corpus of non-translated Russian. Our results indicate lack of direct association between translationese and quality in our data: while our features distinguish translations and non-translations with the near perfect accuracy, the performance of the same algorithm on the quality classes barely exceeds the chance level.

1 Introduction: Aim and Motivation

In the present paper, we test if the linguistic specificity of translations that makes them distinct from non-translations may also reflect their quality. The possible link between translationese and translation quality has been assumed in corpus-based translation studies ever since translationese has become one of the most attractive research topics. At the onset of machine learning approach to translationese detection, Baroni and Bernardini (2006) suggested using machine learning techniques to develop an automatic translationese spotter to be used in translator education. Attempts has been made to correlate translation quality and statistical differences between translations and non-translations in the target language (TL, Scarpa, 2006) and to describe translational tendencies with the view of using them as translation quality assessment tools (Rabadán et al., 2009). Generally, it seems reasonable to posit that the more rigorous the translationese effects, the stronger they signal the low quality of translation. Mostly, the presence of translationese is assumed to affect the fluency of translations, hampering their readability and giving them the distinct flavour of foreignness. While it is true that fluency is one of the traditional aspects of translation quality evaluation, along with pragmatic acceptability and semantic accuracy (as set out in Koponen, 2010; Secara, 2005, for example), it is not clear whether the features that capture translationese can be related to the quality in human translation evaluation. Therefore, we test whether linguistic features responsible for translationese effects are also good indicators of human translation quality as perceived by human experts in real-life educational environment. To the best of our knowledge, the direct application of automatically retrieved translationese features for learning human translation quality has not been attempted before. If successful, this application could be useful for a number of translation technologies, especially those involving automatic quality assessment of both human and machine translation

Temnikova, I, Orăsan, C., Corpas Pastor, G., and Mitkov, R. (eds.) (2019) *Proceedings of the 2nd Workshop on Human-Informed Translation and Interpreting Technology (HiT-IT 2019)*, Varna, Bulgaria, September 5 - 6, pages 47–56.
https://doi.org/10.26615/issn.2683-0078.2019_006

(MT).

We select a range of lexico-grammatical features that have originated in register studies (Biber, 1995; Neumann, 2013) and are known to capture translationese, i.e. to reflect the systemic differences between translated and non-translated texts (see, for example Evert and Neumann, 2017, where they use a similar set to register features to reveal asymmetry in translationese effects for different translation directions in English-German language pair). Importantly, our features are designed as immediately linguistically interpretable as opposed to surface features, such as n-grams and part-of-speech frequencies commonly used in machine translation evaluation, and include manually-checked frequencies of less easily extractable linguistic phenomena such as correlative constructions, nominalisations, by-passives, nouns/ proper names in the function of core verbal arguments, modal predicates, mean dependency distance, etc., along with the more traditional and easily-extractable features like lexical density, frequency of selected parts-of-speech (e.g. subordinating conjunctions and possessive pronouns).

These features are believed to reflect language conventions of the source and target languages (English and Russian in our data) as well as potential 'translationese-prone' areas.

We represent English and Russian texts as feature vectors and use these representations to automatically learn differences between translations/non-translations and high-scoring/low-scoring translations. Assuming that a shift in the translations linguistic properties (away from the target language norm manifested in non-translations) may be related to the translation quality, we use classification techniques to automatically distinguish between good and bad translations. However, we are not only interested in the performance of classifiers, but also in identifying discriminative linguistic features specific either for good or bad translations.

We believe that the findings of this study will contribute to both translation studies and translator training. On the one hand, the knowledge about differences between good and bad translations is important from a didactic point of view, as it delivers information on the potential problems of the novice translators. On the other hand, they provide new insights and new methodological approaches (as our features are automatically retrieved from a corpus) to the area of translation studies and translation technologies.

The remainder of the paper is structured as follows: In Section 2, we report on the related studies and the theoretical background of the paper. Section 3 provides details on our methodology and the resources used. In Section 4 we explore the ability of our features to distinguish between (1) translated and non-translated texts (2) good and bad translations. We report results in terms of accuracy and f-score, and provide a feature analysis. And finally, in Section 5, we conclude and describe the future work.

2 Related Work and Theoretical Background

2.1 Specificity of Translations

Our analyses are based on the studies showing that translations tend to share a set of lexical, syntactic and/ or textual features (e.g. Gellerstam, 1986; Baker, 1995; Teich, 2003). The choice and number of features investigated in translationese studies varies. Corpas Pastor et al. (2008) and Ilisei (2012) use about 20 features to demonstrate translationese effects in professional and student translations from English to Spanish. They used supervised machine learning techniques to distinguish between translated and non-translated texts in this language pair. The authors use two different groups of features – those that grasp general characteristics of texts, e.g. distributions of grammatical words, different part-of-speech classes and the proportion of grammatical words to lexical words, and those that reflect simplification effect (the tendency of translations to be less complex than non-translated texts), such as average sentence length, sentence depth as the parse tree depth, proportion of simple sentences and lexical richness. Our feature set is inspired by the research reported in Evert and Neumann (2017). They adopted 27 features from the feature set developed for the contrastive study in English-German register variation in Neumann (2013) and effectively applied it to the study of translationese effects. This research shows a remarkable similarity between the register features and translationese features: the two sets have a big area of intersection, including, for example, such indicators as sentence length, type-to-token ratio, number of simple sentences, the distributions of some parts-of-speech and function words such as conjunctions, etc. Our own feature set (described in Section 3.2) has considerable extensions and modifications on the one suggested in the works referred above. The feature selection is based on the assumption that the translationese effect is immediately related to quality, and we included the features that are known, or expected, indicators of translationese, which are, incidentally, mostly lexico-grammatical features.

2.2 Translation Features and Quality Estimation

Automatic human translation evaluation is an emerging direction in Natural Language Processing (NLP). For instance, Vela et al. (2014a) and Vela et al. (2014b) used automatic metrics derived from machine translation evaluation and applied them for the evaluation of human translations. They correlated the automatic scores with the human evaluations showing that these automatic metrics should be used with caution. One of the latest work in this strand of research is (Yuan et al., 2016). The authors use easily extractable monolingual features to capture fluency and their bilingual ratios as well as bilingual embeddings features to account for adequacy of content transfer. Their models return the best predictions on the embedding features for both fluency and accuracy. The advantage of using other features such as part-of-speech and dependency frequencies is in their interpretability: the best-performing features selected in their experiments helped the authors to determine grammatical features that are likely to be responsible for lower translation quality scores. They show that human translations typically contain errors beyond the lexical level, to which proximity-based MT evaluation metrics are less sensitive.

The only study that make use of genre features for quality analysis is (Lapshinova-Koltunski and Vela, 2015). However, the authors compare English-German translation (both human and machine) with non-translated German texts that, as the authors claim, represent target language quality conventions. Their main aim is to show that the usage of translation corpora in machine translation should be treated with caution, as human translations do not necessarily correspond to the quality standards that non-translated texts have. Rubino et al. (2016) use features derived from machine translation quality estimation to classify translations and non-translations motivating their work by the fact that automatic distinction between originals and machine translations was shown to correlate with the quality of the machine translated texts (Aharoni et al., 2014). However, their data does not contain human quality evaluation. Translationese as quality indicator was also used by Rabadán et al. (2009) who claims that the smaller the disparity between native and translated usage in the use of particular grammatical structures associated with specific meanings, the higher the translation rates for quality. De Sutter et al. (2017) use a corpus-based statistical approach to measure translation quality (interpreted as target language acceptability) by comparing the features of translated and original texts. They believe that acceptability can be measured as distance to the target language conventions represented in the linguistic behaviour of the professional translators and professional writers. Their analysis is based on the visual estimation of the linguistic homogeneity of professional and original fiction books that are expected to form separate clusters on the Principal Components biplots. The acceptability of student translations is interpreted as the location of a given translation on the plot with regard to these clusters. The PCA-based multivariate analysis was supported by univariate AVOVA tests. The features that were used in this research include a 25 language-independent (overwhelmingly, simple frequencies of parts-of-speech, types, tokens, n-grams, as well as sentence length, TTR, hapax) and 5 language dependent features. The differences observed between professional and student translations are not clear-cut and “only seven features (out of 30) exhibit a significant difference between students and professionals” in their first case study, for example. Their data does not contain manual quality evaluation and it remains unclear how selected linguistic features relate exactly to translation quality. This work is particularly relevant to us, because it is explicitly bringing together translational quality and professionalism.

2.3 Translation Competence

A few other works, like the last one commented above, attempted to capture the specificity of the two translational varieties – the professional and the student translations. If professionalism in translation could be reliably linked to the linguistic properties of translations, (probably, the ones associated with translationese), then professional translations could be used to work around the scarcity and unreliability of the data annotated for translation quality. However, there is hardly any work that has successfully completed this challenging task: professional and learners' translations prove to be difficult to classify. Further product-oriented analyses of professional and student translations that do not exclusively focus on the analysis of errors include works by Nakamura (2007); Bayer-Hohenwarter (2010); Kunilovskaya et al. (2018). The idea to link the level of professional expertise and the performance of a translationese classifier was put to the test in Rubino et al. (2016). They used a range of features to analyse German translations of the two types and non-translated comparable texts in German. Their feature set included features inspired by MT quality estimation (13 surface features such as number of upper-cased letters, and over 700 surprisal and distortion features that were "obtained by computing the negative log probability of a word given its preceding context" based on regular and backward language models). Their result for the binary professional/student translation classification was "barely above the 50% baseline" demonstrating that the MT evaluation features were not helpful for that task. In a similar attempt, Kunilovskaya et al. (2018) used a set of 45 syntactic features (mostly Universal Dependencies relations) to achieve F1 = 0.761, which was lower that their baseline, based on part-of-speech trigrams.

3 Experimental Setup

3.1 Corpus Resources

For our translationese-related analysis, we use a corpus of Russian professional translations to English mass-media texts and a comparable subcorpus of newspaper texts from the Russian National Corpus (RNC, Plungian et al., 2005). Professional translations ('pro') are collected from a range of established electronic media, such as *Nezavisimaya Gazeta* and *InoSMI.RU* or Russian editions of global mass media such as *BBC*, *Forbes* and *National Geographic* (all publications either carry the name of the translator or the endorsement of the translation by the editorial board). Non-translated Russian texts (reference corpus, ref) come from a user-defined subcorpus of the RNC to represent the expected target language norm for the selected register, i.e. the current target language 'textual fit' (Chesterman, 2004). They were sampled on the frame limiting the extracted texts to the type 'article', intended for the large adult non-specialist readership, created after 2003 and marked as neutral of style. For our quality-related analysis, we use the total of 438 student translations from English into Russian labeled for quality in real-life translation competitions, exam or routine classwork settings. All translations were evaluated by the translation experts (either university teachers of translation and/or professional translators), who were asked to rank several translations of the same source text. Though each translation competition and each institution, where translations were graded, had their own descriptions of quality requirements, they were not limiting translation quality to a specific aspect. For the purposes of this research, we relied on the overall agreed judgment of the jury or exam board. For the purposes of this research, we use only 1–3 top ranking translations and/ or translations that received the highest grade and bottom translations and/ or translations that received the lowest grade, which gives us the binary labels 'best' and 'worst'. These translations and their quality labels were extracted from RusLTC (Kutuzov and Kunilovskaya, 2014), a collection of quality-annotated learner translator texts, available online (https://www.rus-ltc.org). The English source texts for both professional and student translations were published in 2001-2016 by well-known English media like *The Guardian*, *The USA Today*, *The New York Times*, *the Economist*, *Popular Mechanics*. All corpus resources used in this research are made comparable in terms of register and are newspaper informational or argumentative texts. The quantitative parameters of the corpus resources used in this research (based on the pre-processed and parsed data) are given in Table 1. We have different number of student translations of the two classes (best, worst), which is also distinct from the number of source texts, because we used several top-ranking translations and in some settings the worst translations were not determined (i.e. the ranking was done only for the top submissions).

Taking into account the small size of our data, we paid attention to its pre-processing to reduce the number of tagging and sentence-splitting errors that may have influence on the feature extraction. First, we normalised spelling and typographic conventions used. Second, we split sen-

		ref	pro	best	worst
EN	**words**	-	458k	49k	
	texts	-	385	98	
RU	**words**	737k	439k	141k	61k
	texts	375	385	305	134

Table 1: Basic statistics on the research corpora

tences with the adjusted NLTK sentence tokeniser, deleted by-lines, dates and short headlines (sentences shorter that 4 tokens, including punctuation) and corrected any sentence boundary errors. Finally, the corpora were tagged with UDpipe 1.2.0 (Straka and Straková, 2017). For each language in this experiments we used the pre-trained model that returned most accurate results for our features and had the highest accuracy for Lemma, Feats and UAS reported at the respective Universal Dependencies (UD) page among the available releases. At the time of writing it is 2.2 for English EWT, and 2.3 for Russian-SynTagRus treebank.

3.2 Features

For our experiments, we use a set of 45 features that include the following types:

- eight morphological forms: two degrees of comparison (`comp, sup`), past tense and passive voice (`pasttense, longpassive, bypassive`), two non-finite forms of verb (`infs, pverbals`), nominalisations (`deverbals`) and finite verbs (`finites`);

- seven morphological categories: pronominal function words (`ppron, demdets, possdet, indef`), adverbial quantifiers (`mquantif`), coordinative and subordinative conjunctions (`cconj, sconj`);

- seven UD relations that are known translationese indicators for the English-Russian translation pair (Kunilovskaya and Kutuzov, 2018). These include adjectival clause, auxiliary, passive voice auxiliary, clausal complement, subject of a passive transformation, asyndeton, a predicative or clausal complement without its own subject (`acl, aux, aux:pass, ccomp, nsubj:pass, parataxis, xcomp`).

- three syntactic functions in addition to UD relations: various PoS in attributive function (`attrib`), copula verbs (`copula`), nouns or proper names used in the functions of core verbal argument (subject, direct or indirect object) to the total number of these relations (`nnargs`);

- nine syntactic features that have to do with the sentence type and structure: simple sentences (`simple`), number of clauses per sentence (`numcls`), sentence length (`sentlength`), negative sentences (`neg`), types of clauses – relative (`relativ`) and pied-piped subtype (`pied`), correlative constructions (`correl`), modal predicates (`mpred`), adverbial clause introduced by a pronominal ADV(`whconj`);

- two graph-based features: mean hierarchical distance and mean dependency distance (`mhd, mdd`) (Jing and Liu, 2015);

- five list-based features for semantic types of discourse markers (`addit, advers, caus, tempseq, epist`) and the discourse marker *but*[1] (`but`). The approach to classification roughly follows (Halliday and Hasan, 1976; Biber et al., 1999; Fraser, 2006). The search lists were initially produced independently from grammar reference books, dictionaries of function words and relevant research papers and then verified for comparability and consistency;

- two overall text measures of lexical density and variety (`lexdens, lexTTR`).

Special effort was made to keep our feature set cross-linguistically comparable. The rationale behind this decision is an attempt to reveal the most notorious effect in translation, namely, 'shining-through', the translational tendency to reproduce source language patterns and frequencies rather than follow the target language conventions. This form of translationese can be established by comparing the distributions of a feature values across three corpora: non-translations in the source language (SL), non-translations (or reference) in the TL and in the translated texts in the TL. We use several norms to make features comparable across different-size corpora, depending on the nature of the feature. Most of the features, including all

[1] If not followed by 'also' and not in the absolute sentence end.

types of discourse markers, negative particles, passives, relative clauses, are normalised to the number of sentences (30 features). Such features as personal, possessive pronouns and other noun substitutes, nouns, adverbial quantifiers, determiners are normalised to the running words (6 features). Counts for syntactic relations are represented as probabilities, normalised to the number of sentences (7 features). Some features use their own normalisation basis: comparative and superlative degrees are normalised to the total number of adjectives and adverbs, nouns in the functions of subject, object or indirect object are normalised to the total number of these roles in the text.

3.3 Methodology

We extract the instances of the features from our corpus relying on the automatically annotated structures (parts-of-speech, dependency relations, etc.). The accuracy of feature extraction is therefore largely related to the accuracy of the automatic annotation. However, care has been taken to filter out noise by using empirically-motivated lists of the closed sets of function words and typical annotation errors where possible. Each text in the data is represented as a feature vector of measures for a range of linguistic properties as described in 3.2.

For both tasks – (1) the analysis of the differences between translated and non-translated texts and (2) the comparison of the highest-ranking and lowest-ranking translations, we model the difference between our binary text classes using machine learning techniques. The experiments are arranged as text classification tasks, where we determine the utility of our features based on the performance of the classifier. For the consideration of space, we report the results of a Support Vector Machine (SVM) algorithm with the default sklearn hyper parameters only. To account for the generalization error of the classifier, we cross-validate over 10 folds. The results of the same learner on the full feature set are compared to the results on the most informative features only to reveal the comparative usefulness of our handcrafted features for each task. Below we report the results for the 15 best features selected with Recursive Feature Elimination (RFE) method, which seems preferable to the standard ANOVA-based SelectKBest, because some of our features do not comply with the normal distribution assumption made by ANOVA. Besides, we use Principal Component Analysis (PCA) to visualise the distinctions between our classes, given our features.

In the first task, we automatically distinguish comparable Russian non-translations from professional and student translations. In the second task, we use the same algorithm and the same features to learn the difference between good and bad translations. The comparative outcome of this two-step methodology indicates whether the features described in 3.2 capture translationese, whether they correlate with the human evaluation of human translation quality, and whether there is an association between the two. Moreover, we analyse which features are most informative in the two classification tasks and intersect the resulting feature lists.

4 Results and their Interpretation

4.1 Translationese

As seen in Figure 1 illustrating the results of PCA, our features are good indicators of translationese: we get very similar, consistent results on the differentiation between the non-translations in our data and the two translational corpora that come from different sources and, in fact, represent two socio-linguistic translational varieties (student and professional translations).

These visual impressions are corroborated by the results of the automatic classification. Table 2 show that this feature set allows us to predict translations of any type with the accuracy of 92-94%.

	precision	**recall**	**f1-score**
pro	0.91	0.94	0.93
ref	0.94	0.91	0.92
macro avg	0.92	0.92	0.92
stu	0.93	0.95	0.94
ref	0.94	0.92	0.93
macro avg	0.94	0.94	0.94

Table 2: Cross-validated classification between translations and non-translations on the full feature set

As a sanity check measure, we ran a dummy classifier that randomly allocates labels with respect to the training set's class distribution to get the expected overall accuracy of 48%. Most informative features contributing to this distinction (as selected by RFE wrapped around a Random Forest algorithm) include `possdet, whconj,`

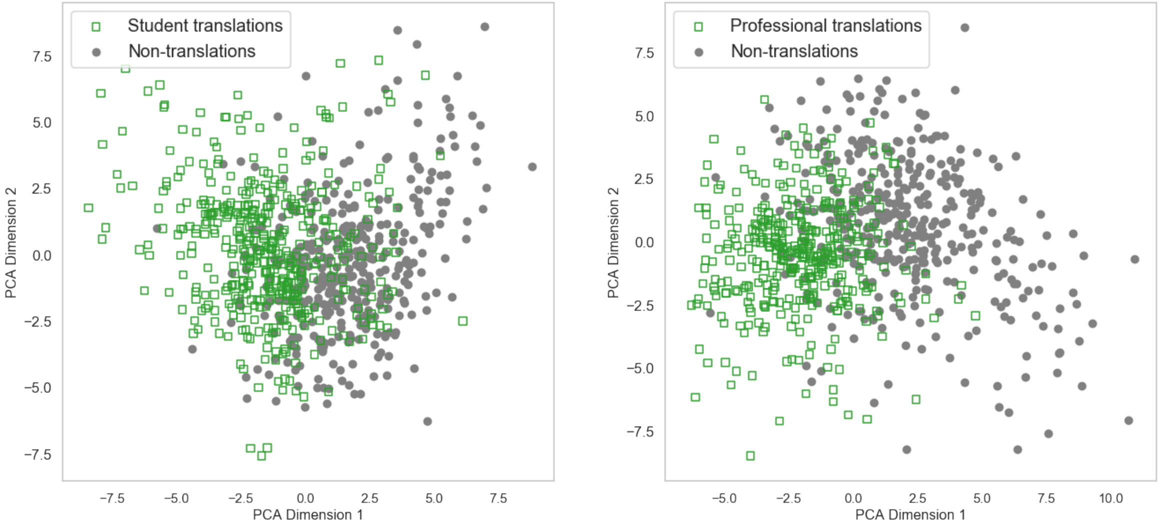

Figure 1: Student and professional vs. non-translations in Russian

`relativ, correl, lexdens, lexTTR, finites, deverbals, sconj, but, comp, numcls, simple, nnargs, ccomp`. It is the stable best indicators of translationese: 2/3 of this list is reproducible on the both translational collections, and the classification results on just these features are only 3% inferior to the whole 45-feature set.

4.2 Quality

Using the same feature set, we analyse differences between the top-scoring and lowest-scoring translations labelled as 'good' and 'bad' in our data. As seen from Figure 2 that plots the values for our data points on the first two dimensions from PCA (the x- and y-axis, respectively), the best and the worst translations are evenly scattered in the two-dimensional space and, unlike the previous experiment, no groupings are visible.

The cross-validated SVM classifier on the full feature set for good/bad translations returns the macro-averaged F1-measure of 0.64 (Table 3). The overall accuracy of this classification is 68%. Interestingly, good translations can be more easily modelled than the bad ones (76% vs. 51% respectively). This contradicts expectations from the teaching practice where examiners commonly better agree on what is a bad translation. But given that bad translations are a minority class in our classification and that the employed feature set performs worse than a dummy classifier which achieves 73% accuracy, these observations are unreliable anyway. The result on the 20 RFE features is the same as on the full feature set of 45, but worse than that returned by the dummy classifier.

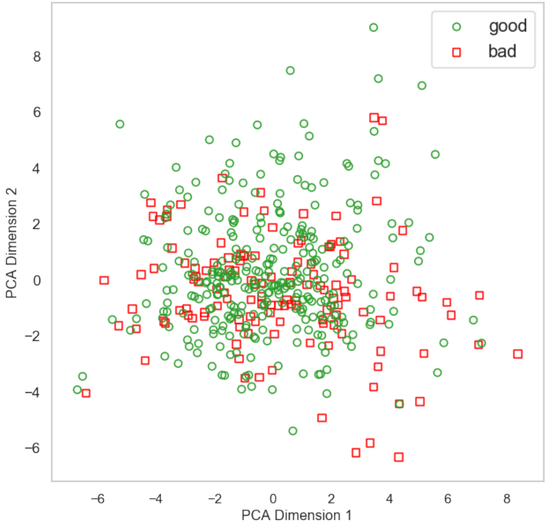

Figure 2: Best vs. worst translations

	precision	recall	f1-score
bad	0.48	0.55	0.51
good	0.79	0.74	0.76
macro avg	0.63	0.64	0.64

Table 3: Results for good/bad classification

If we attempt the classification on the 15 best translationese indicators established in the previous step of this research, we would see the overall classification results deteriorate to F1=0.56, while the results for the minority class ('bad') plummet to F1=0.36.

Even though the classification result can hardly be found reliable, we calculated the features that statistically return the best differentiation between the labeled classes according to ANOVA. They include `copula, finites, pasttense, infs, relativ, lexdens, addit, ccomp, but, sconj, nnargs, acl, advers, ppron, sentlength`. The intersection with the 15 top translationese indicators is limited to the six list items: `finites, lexdens, but, relativ, nnargs, sconj, ccomp`.

One of the major motivation behind this research was to reveal the existence and extent of features responsible for one distinct form of translationese, namely, shining-through. We visualise the difference (distance) between good and bad translations with a kernel density estimation (KDE) plot provided in Figure 3. This plot demonstrates how well the values learnt on one of the PCA dimensions separate the text classes in our experiment. In this way, we are able to observe the extent of the shining through effects in our data: while it is clear that all translations are located in the gap between the source and the target language, this form of translationese does not differentiate translations of different quality. If shining through features were useful in discerning bad translations (as we expected), the red line should have been more shifted towards the yellow dashed line of the source language. Needless to say, the professional translations demonstrate a similar shining through effect, which we do not illustrate here for brevity.

5 Conclusion

In the present paper, we analyzed if morphosyntactic features used in register studies and translationese studies are also useful for the analysis of quality in translation. It is often assumed that any differences of translations from non-translations may affect the fluency of translations. If so, automatically extracted translationese features can also be used for human translation evaluation, which saves time and effort of manual annotation for quality.

We tested this on a dataset containing English-Russian translations that were manually evaluated for quality. The results of our analysis show that features that are good for predicting translationese, i.e. separating translations from the comparable non-translations, are not necessarily good in predicting translation quality, at least for the data at hand. We have to admit that these results do not align well with our expectations. One explanation is that we relied on the morphology and syntax for capturing translationese, while the most immediately perceptible lexical level remained unaccounted for. Another reason for the lack of correlation between the quality labels and the fluency (understood here as deviations from TL morphosyntactic patterns) is that quality is not entirely about fluency, of course. The quality labels in our data must reflect semantic faithfulness and pragmatic acceptability of translations as well. If anything, our results support the original interpretation of translationese as inherent properties of translations exempt from the value judgment: translationese is not the result of poor translation, but rather a statistical phenomenon: various features distribute differently in originals than in translations (Gellerstam, 1986).

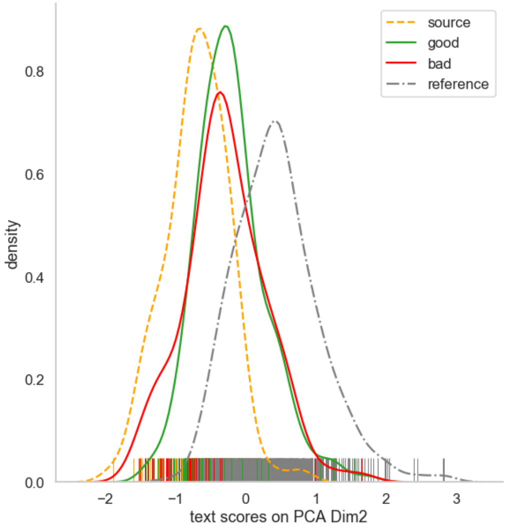

Figure 3: Good and bad translations vs. non-translations in the source and the target languages

To our knowledge, there are no further studies pursuing direct application of translationese features for learning human translation quality. In (De Sutter et al., 2017), the authors tried to automatically assess translation quality of student translations measuring their deviation from the "normal" texts represented by professional translations and non-translated texts in a target language. Although they were able to show that student translations differ from both comparable originals and professional translations, it is not clear if these differences were encountered due to other influencing factors, as their data does not contain any manual evaluation. Besides that, they were not able to find out why certain linguistic features were indicators of deviant student translation behaviour in a given setting.

Similarly, we show that translationese, at least the features used in our analysis, are not necessarily good indicators of translation quality. We believe that these results provide valuable insights for both translation studies and translation technologies, especially those involving quality estimation issues.

Acknowledgments

This work is mostly produced in the University of Tyumen and is supported in part by a grant (Reference No. 17-06-00107) from the Russian Foundation for Basic Research.

References

Roee Aharoni, Moshe Koppel, and Yoav Goldberg. 2014. Automatic detection of machine translated text and translation quality estimation. In *Proceedings of ACL*, pages 289–295.

Mona Baker. 1993. Corpus linguistics and translation studies: Implications and applications. In G. Francis Baker M. and E. Tognini-Bonelli, editors, *Text and Technology: in Honour of John Sinclair*, pages 233–250. Benjamins, Amsterdam.

Mona Baker. 1995. Corpora in translation studies: An overview and some suggestions for future research. *Target*, 7(2):223–243.

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

Gerrit Bayer-Hohenwarter. 2010. Comparing translational creativity scores of students and professionals: flexible problem-solving and/or fluent routine behaviour? In S. Göpferich, F. Alves, and I. Mees, editors, *New Approaches in Translation Process Research*, Copenhagen studies in language, pages 83–111. Samfundslitteratur.

Douglas Biber. 1995. *Dimensions of Register Variation: A Cross-Linguistic Comparison*. Cambridge University Press.

Douglas Biber, Susan Conrad, Edward Finegan, Stig Johansson, Geoffrey Leech, Susan Conrad, Edward Finegan, and Randolph Quirk. 1999. *Longman grammar of spoken and written English*, volume 2. MIT Press.

Andrew Chesterman. 2004. Hypotheses about translation universals. *Claims, Changes and Challenges in Translation Studies*, pages 1–14.

Gloria Corpas Pastor, Ruslan Mitkov, Naveed Afzal, and Lisette Garcia-Moya. 2008. Translation universals: do they exist? a corpus-based and nlp approach to convergence. In *Proceedings of the LREC-2008 Workshop on Building and Using Comparable Corpora*, pages 1–7.

Gert De Sutter, Bert Cappelle, Orphée De Clercq, Rudy Loock, and Koen Plevoets. 2017. Towards a corpus-based, statistical approach to translation quality: Measuring and visualizing linguistic deviance in student translations. *Linguistica Antverpiensia, New Series–Themes in Translation Studies*, 16.

Stefan Evert and Stella Neumann. 2017. The impact of translation direction on characteristics of translated texts : A multivariate analysis for English and German. *Empirical Translation Studies: New Methodological and Theoretical Traditions*, 300:47.

Bruce Fraser. 2006. Towards a Theory of Discourse Markers. *Approaches to discourse particles*, 1:189–204.

Martin Gellerstam. 1986. Translationese in Swedish novels translated from English. In L. Wollin and H. Lindquist, editors, *Translation Studies in Scandinavia*, pages 88–95. CWK Gleerup, Lund.

M A K Halliday and Ruqaiya Hasan. 1976. *Cohesion in English*. Equinox.

Iustina Ilisei. 2012. *A machine learning approach to the identification of translational language: an inquiry into translationese*. Doctoral thesis, University of Wolverhampton.

Iustina Ilisei, Diana Inkpen, Gloria Corpas Pastor, and Ruslan Mitkov. 2010. Identification of translationese: a supervised learning approach. In *Proceedings of CICLing-2010*, volume 6008 of *LNCS*, pages 503–511, Springer, Heidelberg.

Yingqi Jing and Haitao Liu. 2015. Mean Hierarchical Distance Augmenting Mean Dependency Distance. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 161–170.

Maarit Koponen. 2010. Assessing Machine Translation Quality with Error Analysis. In *Electronic proceedings of the VIII KäTu symposium on translation and interpreting studies*, volume 4, pages 1–12.

Maria Kunilovskaya and Andrey Kutuzov. 2018. Universal Dependencies-based syntactic features in detecting human translation varieties. *Proceedings ofthe 16th International Workshop on Treebanks and Linguistic Theories (TLT16)*, pages 27–36.

Maria Kunilovskaya, Natalia Morgoun, and Alexey Pariy. 2018. Learner vs. professional translations into Russian: Lexical profiles. *Translation & Interpreting*, 10.

Andrey Kutuzov and Maria Kunilovskaya. 2014. Russian learner translator corpus. In Petr Sojka, Aleš Horák, Ivan Kopeček, and Karel Pala, editors, *Text, Speech and Dialogue*, volume 8655 of *Lecture Notes in Computer Science*, pages 315–323. Springer International Publishing.

Ekaterina Lapshinova-Koltunski and Mihaela Vela. 2015. Measuring 'registerness' in human and machine translation: A text classification approach. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 122–131, Lisbon, Portugal. Association for Computational Linguistics.

Sachiko Nakamura. 2007. Comparison of features of texts translated by professional and learner translators. In *Proceedings of the 4th Corpus Linguistics conference*, University of Birmingham.

Stella Neumann. 2013. *Contrastive register variation. A quantitative approach to the comparison of English and German.* Mouton de Gruyter, Berlin, Boston.

Vladimir Plungian, Tatyana Reznikova, and Dmitri Sitchinava. 2005. Russian National Corpus: General description [Nacional'nyj korpus russkogo jazyka: obshhaja harakteristika]. *Scientific and technical information. Series 2: Information processes and systems*, 3:9–13.

Rosa Rabadán, Belén Labrador, and Noelia Ramón. 2009. Corpus-based contrastive analysis and translation universals A tool for translation quality assessment. *Babel*, 55(4):303–328.

Raphael Rubino, Ekaterina Lapshinova-Koltunski, and Josef van Genabith. 2016. Information density and quality estimation features as translationese indicators for human translation classification. In *Proceedings of NAACL HT 2006*, pages 960–970, San Diego, California.

Federica Scarpa. 2006. Corpus-based quality-assessment of specialist translation: A study using parallel and comparable corpora in English and Italian. In Maurizio Gotti and Susan Šarčevic, editors, *Insights into specialized translation*, volume 46 of *Linguistic Insights / Studies in Language and Communication*, pages 155–172. Peter Lang, Bern.

Alina Secara. 2005. Translation Evaluation - a State of the Art Survey. *Proceedings of the eCoLoRe/MeLLANGE Workshop, Leeds*, pages 39–44.

Milan Straka and Jana Straková. 2017. Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99.

Elke Teich. 2003. *Cross-Linguistic Variation in System and Text. A Methodology for the Investigation of Translations and Comparable Texts*. Mouton de Gruyter, Berlin.

Mihaela Vela, Anne-Kathrin Schumann, and Andrea Wurm. 2014a. Beyond Linguistic Equivalence. An Empirical Study of Translation Evaluation in a Translation Learner Corpus. In *Proceedings of the EACL 2014 Workshop on Humans and Computer-assisted Translation*, pages 47–56, Gothenburg, Sweden. Association for Computational Linguistics.

Mihaela Vela, Anne-Kathrin Schumann, and Andrea Wurm. 2014b. Human Translation Evaluation and its Coverage by Automatic Scores. In *Proceedings of MTE Workshop at LREC 2014*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Yu Yuan, Serge Sharoff, and Bogdan Babych. 2016. MoBiL: A Hybrid Feature Set for Automatic Human Translation Quality Assessment. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*.

Frederico Zanettin. 2013. Corpus Methods for Descriptive Translation Studies. *Procedia - Social and Behavioral Sciences*, 95:20–32.