Fiction in Russian Translation: A Translationese Study

Maria Kunilovskaya[1], Ekaterina Lapshinova-Koltunski[2], and Ruslan Mitkov[3]

[1,3]University of Wolverhampton
[2]Saarland University
[1,3]maria.kunilovskaya@wlv.ac.uk, r.mitkov@wlv.ac.uk
[2]e.lapshinova@mx.uni-saarland.de

Abstract

This paper presents a translationese study based on the parallel data from the Russian National Corpus (RNC). We explored differences between literary texts originally authored in Russian and fiction translated into Russian from 11 languages. The texts are represented with frequency-based features that capture structural and lexical properties of language. Binary classification results indicate that literary translations can be distinguished from non-translations with an accuracy ranging from 82 to 92% depending on the source language and feature set. Multiclass classification confirms that translations from distant languages are more distinct from non-translations than translations from languages that are typologically close to Russian. It also demonstrates that translations from same-family source languages share translationese properties. Structural features return more consistent results than features relying on external resources and capturing lexical properties of texts in both translationese detection and source language identification tasks.

1 Introduction

This paper reports results of a translationese study based on literary texts translated from 11 source languages into Russian and included into the Russian National Corpus[1]. Translationese is understood as specific linguistic properties of translations distinguishing them from non-translated language. Most existing investigations into the properties of translations have focused on text types other than fiction and, to the best of our knowledge, there were no large-scale investigations into translational Russian, especially based on literary texts. The research applications of the parallel multilingual corpus of translations into Russian are limited to constrastive studies such as (Dobrovol'skij and Pöppel, 2017) and translation studies such as (Krasnopeyeva, 2016), and deal with a few individual items or constructions. We explore linguistic differences between fiction translated into Russian and originally-authored literature in Russian with a focus on the impact of the source language (SL).

The comparison between translated and non-translated Russian is based on an extended feature set capturing structural and abstract lexical parameters of texts, including collocational properties. This investigation mostly relies on the text classification approach and univariate statistical analyses following a methodology established in computational studies of translationese (Baroni and Bernardini, 2006; Volansky et al., 2015; Evert and Neumann, 2017, amongst others). Translationese indicators that cut across all translated subcorpora (if any) or are specific to a particular language pair are identified through feature selection.

It does not come as a surprise that researchers are wary of using fiction for translationese studies. Fiction can be less homogeneous register because each literary work and author might display unique conceptual and linguistic properties reflecting a particular manner of artistic expression and idiolect. With literary translation established as a form of creative endeavour of its own standing and with the covert 'domesticating' translation paradigm accepted as a professional standard, little difference between translations and non-translations is to be expected in this register. This study is based on a carefully balanced and representative sample of literary texts from each language pair designed to reduce possible impact of author/translator idiolects on the results.

We aim to establish if and how Russian non-translated literary texts differ from literary translations from a variety of SLs (our first research

[1]See http://ruscorpora.ru/

Proceedings of Recent Advances in Natural Language Processing, pages 734–743
Sep 1–3, 2021.
https://doi.org/10.26615/978-954-452-072-4_084

question, **RQ1**).

Following a recent trend in translationese studies (Dutta Chowdhury et al., 2020; Bjerva et al., 2019; Rabinovich et al., 2017), we are interested in exploring the link between the amount of translationese and the SL involved. Translations from more distant languages were shown to return better classification results indicating greater differences from non-translations. Can we corroborate this finding on our data and features (our second research question, **RQ2**)?

Our results are relevant to typological and contrastive studies that use parallel data to draw conclusions about the properties of target language (TL) because we are highlighting that translations constitute a TL variety, significantly diverging from comparable non-translations. This study also contributes to a direction in translationese studies that seeks to establish links between SL-TL cross-linguistic distance and the amount of observed translationese as well as to the task of document-level SL detection.

2 Related Work

Translations vs non-translations Our work is related to studies showing that translated texts share linguistic features that make them distinct from non-translations (Gellerstam, 1986; Baker, 1993). These features have been useful in automatic classification of translated and non-translated texts (Baroni and Bernardini, 2006; Volansky et al., 2015; Rubino et al., 2016; Kunilovskaya and Lapshinova-Koltunski, 2020). Some studies show that combinations of features perform better in machine learning settings for translationese-related tasks (Lynch and Vogel, 2012; Evert and Neumann, 2017; Sominsky and Wintner, 2019). This paper compares the effectiveness of two feature sets: morphosyntactic and lexical indicators. The former are count-based features extracted directly from automatically annotated research data. The latter are estimated using language models trained on a large, register-comparable TL corpus. They capture dissimilarity of translations with regard to the TL norm manifested in this additional TL resource.

Translationese in literary texts There are several corpus-based studies of literary translations. Most of them explore single features: passive constructions (Kolehmainen and Riionheimo, 2016), *that*-complementiser (Olohan, 2001), non-finite constructions, phrasal verbs, connectives (Kunilovskaya, 2017), keywords (Puurtinen, 2003), etc.

The only two works known to us that use machine learning to study translationese in literary texts include (Popescu, 2011) and (Lynch and Vogel, 2012). The first study explored the effectiveness of character 5-grams comparing originally-written English literature with translations from French and German. The second study is more similar to the design proposed in this study. It reports the results of SL detection based on English literary translations from Russian, German and French. A Support Vector Machine (SVM) 4-class classifier achieves the accuracy of 80% in the train-test scenario using a combined set of 50 features. The authors attempt the analysis of some of the top-ranking features, trying to link them to the SLs. In our work, we also employ multiclass classification to identify best translationese predictors. However, we rely on very different features and experiment on a twice bigger data from a wider range of SLs.

Tracing source languages Source language detection is based on the assumption that translations tend to retain features of the SLs (shining through effect, Teich, 2003). Rabinovich et al. (2017) showed that translations carried enough signal from SLs to restore the phylogenetic language tree. The authors experimented on English translations of the European parliamentary speeches in 17 SLs representing three language families. They noticed that misclassified instances were frequently assigned to genetically related languages. Translations from SLs with isomorphic structures displayed a tendency to share more translationese features. Another study by Bjerva et al. (2019) used three levels of syntactic abstraction to explore genetic, geographical and structural distances between SLs. In their results, structural similarities were a better predictor of similarities between languages than genetic ones. Dutta Chowdhury et al. (2020) used isomorphism between embedding spaces, hypothesising that the more isomorphism was detected between translations into English and non-translated English, the closer the source and target languages. They learned delexicalised multi-view representations – embeddings based on parts-of-speech (PoS) tags, lexical semantic tags, and conceptual-semantic tags from WordNet. However, these studies are based on EuroParl, a corpus of parliamentary speeches, which might be more homogeneous in style, more conventionalised as a

text type and in terms of translation strategies than literary texts focused in this paper.

The relation between predictability of translated texts and the divergence between the TL and the SL in terms of morpho-syntax were analysed in a recent study by Nikolaev et al. (2020). The authors used Parallel Universal Dependencies to analyse 1000 sentences from the news domain and Wikipedia translated from English into eight languages. The results showed that translations from similar and distant languages were both predictable, but in different ways: structurally-similar SLs favoured the use of a narrower range of syntactic patterns limited to those shared by two languages, which constituted one type of translational specificity. In translations from highly-divergent languages, however, translators tended to produce non-idiomatic renditions, that were not recognised by models trained on the TL. Sominsky and Wintner (2019) also used the SL signal to detect translation direction. They found that the more distant were the source and the target languages, the higher the SL-detection results.

3 Methodology

3.1 Research Corpus

Our data comes from the parallel component of the RNC. It is a bidirectional corpus, which contains translations from and into Russian for 19 language pairs. Fiction is by far the most represented register, however Russian translations of newspaper texts from some SLs are also included. At the time of writing, parallel RNC includes Russian translations of fiction from 19 SLs, with the overall size of this translational material estimated at 36.6 million tokens. This corpus was sampled to include all language pairs with Russian as the TL that have at least five documents of lengths over 20k tokens produced by a unique combination of author and translator. The document size limit is intended to exclude short stories and retain only novellas and novels. In the absence of a genre annotation, this restriction maintains some comparability of our documents with regard to the type of literary work. The author/translator condition minimises the impact of individual writing style on our results. Another data selection constraint ensures that translations were produced within a time span of 100 years (1925–2020). We excluded document pairs where the author's mother tongue was not the respective SL (e.g. Nabokov) and where the translations were done by the author (e.g. Vasil Bykov). This sampling frame leaves us with a corpus of 210 document pairs in 11 parallel subcorpora. At preprocessing stage, we discarded all sentence pairs with empty source or target as well as by-lines and headings.

To build a comparable collection of non-translated Russian fiction, we used the same sampling frame on the monolingual part of the RNC. We retained only the largest work by every author, deleted the works in Russian by bilingual authors (e.g. Nabokov) and novels explicitly marked as translations. These selection criteria yielded a corpus counting 439 documents (longer than 20K tokens), 33.8M tokens (3.2M sentences) in total. Details on the distribution of document lengths in the samples for each SL and Russian reference sample are presented in Figure 1. This material was used for chunking (see below).

The SLs in the resulting collection are distributed among four language families in our collection: Romance (French, Spanish), Germanic (Swedish, English, German), Balto-Slavic (Baltic: Latvian, Slavic: Polish, Belarusian, Ukrainian, Bulgarian), Uralic (Finnish) based on linguistically motivated phylogenetic language tree (see Serva and Petroni, 2008). In total, we have 859 documents in our data, including parallel and monolingual components. To balance the data in terms of document size, we randomly selected 10 books with unique author-translator combinations. For Bulgarian, Spanish, French, Finnish and Latvian we had to do with fewer books to meet this restriction. After that, each literary work was chunked into portions of 100 consecutive sentences and 10 random chunks were extracted. This collection of text chunks was used in the experiments below.

3.2 Features for classification

The key component of this methodology is the features. We classify them into structural and abstract lexical features.

Structural features. The first subset includes 45 features extracted from Universal Dependencies (UD) annotations of the data (UD features). The values for UD features are normalised frequencies of various UD tags and their combinations, reflecting the morphological and syntactic structure of language. They are selected to capture the differences in the linguistic make-up of translations demonstrated in translationese studies for other lan-

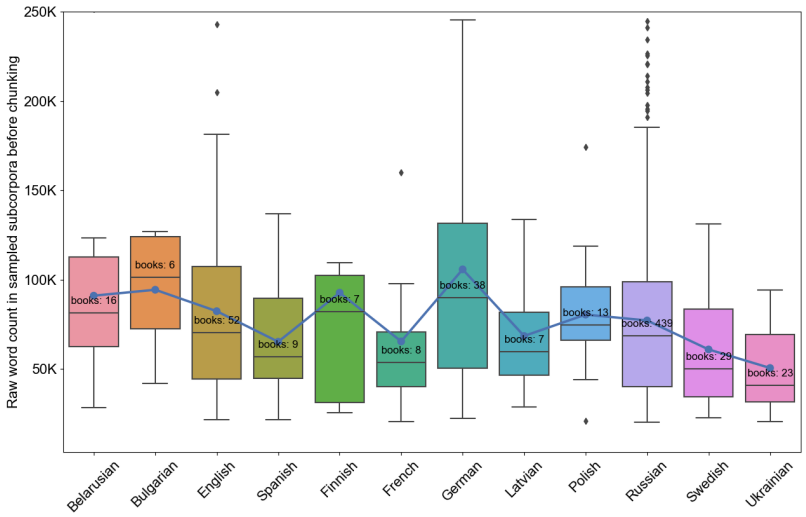

Figure 1: Document sizes and number of books by SL in the filtered research corpus

guage pairs and anticipated in out-of-English Russian translations in the practical translation textbooks based on the typological differences of English and Russian. It has been shown that translations usually have lower type-to-token ratio, higher sentence length and greater number of connectives in a number of language pairs. It is also expected that Russian translations from English would have higher frequencies of modal predicates, analytical passive forms, inflated frequencies of several types of pronouns due to typological differences. The values for most features are cumulative frequencies of all lemmas that belong to a word class or all forms that received a specific tag. We summarise the types of structural features in Table 2 in Appendix[2].

The normalisation basis varies depending on the type of item, following the motivation in Evert and Neumann (2017): it is total text tokens for word classes; number of sentences for conjunctions, modal predicates; total verbs for verb forms; total number of dependencies for the select types of dependencies. The values for discourse markers features are normalised cumulative frequencies of four semantic types of connectives and of epistemic markers (e.g. *of course, probably, actually*), extracted based on pre-defined lists of lemmas (183 items in total for connectives and 86 items for epistemic markers). The lists are informed by Russian grammars and special linguistic dictionaries. Though most items on the lists are set phrases, we allowed for possible lexical and structural variability during extraction. We also used positional heuristics and punctuation to disambiguate our items. The output of the extraction procedure was manually checked to exclude greedy matching.

Abstract lexical features. The second subset, counting 23 features, requires language models learnt from a separate (bigger) corpus resource of original Russian literature (LM features). We used the corpus described in Section 3.1, excluding the 10 random books used to get 100 chunks of non-translated Russian subcorpus. We used a 3-gram language model learnt on this corpus with KenLM library (Heafield, 2011) to generate average sentence perplexities and their standard deviation for each text chunk in our data. We hypothesise that this model will return higher perplexities depending on how unusual the sequences of lexical items are in the translated language. N-grams frequency lists (of orders 1, 2 and 3) from the same corpus were used to calculate the ratios of lemmas that belonged to the highest- and lowest-frequency quartiles of this list for each order of the n-grams cumulatively. Ratios of out-of-vocabulary (OOV) items were used as separate features for each n-gram order. These features were supposed to capture the overuse of the TL high-frequency items and, possibly, a higher ratio of OOV items. These features were inspired by feature-based quality estimation approaches used for machine trans-

[2]The extraction code is available from https://github.com/kunilovskaya/translationese45

lation (Specia et al., 2015). We also experimented with 12 collocational features that were assumed to capture various aspects of co-occurrence patterns in the data. To define these features we relied on the concept of collgrams, explained by Bestgen and Granger (2014) as n-grams with an association score above an arbitrary threshold. For association measures we used normalised pointwise mutual information (NPMI) and t-score to detect collgrams composed of less-frequent and more-frequent words respectively. For each association measure we trained a bigram model on the large reference corpus described above with the Phrases module from the Gensim library (Řehůřek and Sojka, 2010). Then, the model was applied to the chunks of text in the test corpus. In this approach the collgrams were detected and the average association scores across all collgrams in each chunk were produced based on the frequency statistics in the reference corpus. T-score measure was calculated using the formula from (Gries, 2010). For NMPI we relied on the Gensim inbuilt scorer (Bouma, 2009). In both measures 0 means independence of bigram components; NPMI lies in the [-1: 1] range, while t-score bounds were experimentally established within [-11: 9] scope. To train the models, we set the association score threshold to the lower bound of each metric. The bigram frequency threshold was set to 1 to score all bigrams in the 33-million reference corpus. While learning the phraser model, we allowed for intervening words from the functional word classes to access items like альфа::и::омега *(alpha::and::omega)*, пакт::о::ненападение *(non::aggression::pact)*. In calculating feature values we relied on bigram/collgram types, not tokens. To sum up, the collocational features for each association measure extracted from each chunk in the research corpus include: (1) ratio of highly-associated collgrams to all bigrams (the cut-off for high association was set to recommended NPMI > 0.5 and t-score > 6); (2) ratio of negatively-associated collgrams to all bigrams; (3) ratio of all detected collgrams with the score > 0 to the total word count; (4) ratio of bigrams absent in the model to all bigrams in the test corpora; (5) average association score for all detected collgrams with the association score > 0; (6) standard deviation for the association scores in each text chunk.

We expect that translations would have a higher ratio of t-score-based highly-associated bigrams, and lower ratio of NPMI-based highly-associated bigrams than in the comparable subcorpus of non-translations. This hypothesis is based on the known properties of translation to prefer high-frequency items and on the known properties of the association measures: MI is known to favour sequences made of low-frequency items, while t-score assigns higher scores to high-frequency items (Gries, 2010). The ratio of negatively associated bigrams and the ratio of bigrams not seen in the reference is aimed to capture less usual sequences which can be a sign of shining though or errors, including acceptability in the register (e.g. тяжелая критика, крепкая основа).

Finally, we calculate the average association score for each text chunk to reflect general 'collacatedness' of translations and non-translations and standard deviation across all chunks in each source-language subcorpus. All lexical features were produced on lemmatised corpora, where proper names and their sequences were replaced with PROPN and all numbers were represented as XXX (e.g. Борис Николаевич Юрьев -> *PROP*, 1984 -> XXXX). We also deleted all punctuation, except end-of-sentence marks.

The features were extracted into a table containing 1060 rows representing all text chunks in our experiments, labelled with SL, including 'ru' for chunks selected from the reference corpus and 68 features, including 45 UD-based (inc. list-based features), 11 ngram-based and 12 collgram-based ones. The features in the last two groups were extracted with reference to several LMs learnt from a reference corpus of non-translated fiction. The input feature values were z-transformed with the scikit-learn Standard Scaler.

4 Results and Discussion

To find out whether our features capture any differences between translations and non-translations in SL subcorpora and whether the scale of these differences is traceable to the typological group of the SL, we ran 11 binary classifications and a multiclass classification. In all experiments we used Support Vector Machine (SVM) algorithm with the default scikit-learn settings (C=1.0, kernel='rbf', degree=3, gamma='scale'). For the five subcorpora where we have fewer observations than in the reference sample, we used class_weight='balanced' option.

Table 1 presents the accuracy (acc) and macro

		all 68		45 UD		11 ngram		12 collgram		baseline	
SL	docs	acc in%	F1	acc in%	F1	acc in%	F1	acc in%	F1	acc in%	F1
Distant languages (Germanic, Romance and Finnish)											
en	200	89.0	.89	87.5	.87	71.0	.70	71.0	.70	50.0	.33
es	190	86.3	.86	84.7	.84	65.3	.64	63.7	.63	63.2	.63
ge	200	91.0	.91	87.0	.87	70.5	.70	67.5	.67	50.0	.33
fi	170	87.0	.86	82.9	.82	71.8	.69	64.7	.62	47.0	.46
fr	170	82.9	.92	**91.8**	.91	61.2	.57	65.3	.63	47.0	.46
sv	200	90.5	.90	91.0	.91	70.5	.70	77.0	.77	50.0	.33
Close Slavic languages, inc. Balto-Slavic:											
be	200	83.0	.83	82.5	.82	72.5	.72	67.0	.66	50.0	.33
bg	160	88.1	.86	85.0	.83	68.8	.59	68.7	.59	50.0	.33
lv	170	82.9	.82	79.4	.79	65.3	.63	67.1	.64	47.0	.46
pl	200	82.0	.82	79.5	.79	57.0	.56	64.5	.64	50.0	.33
uk	200	85.5	.85	83.0	.83	74.5	.74	72.0	.72	50.0	.33

Table 1: Binary classification results by SL group and feature set; be=Belarusian, bg=Bulgarian, en=English, es=Spanish, fi=Finnish, fr=French, ge=German, lv=Latvian, pl=Polish, sv=Swedish, uk=Ukrainian.

F1-score (F1) results of the binary classifications for each translational subcorpus (represented by SL) against non-translations. The SLs are presented in two groups: Typologically distant and close languages (in relation to Russian). We also group the scores according to the feature sets and combinations: all 68 features, 45 UD features (structural features), 11 n-gram and 12 collgram (both abstract lexical features). The last two columns show the results expected by chance. This simple baseline was implemented as a random classifier which predicts classes with respect to the distribution of instances across classes. We report the scores for the 10-fold cross-validation scenario.

RQ1 Overall, we are able to detect differences between translated and non-translated fiction in Russian. Combined features and each feature set individually performed better than the chance-level baseline, except collocational features for translated Spanish, which were on par with the dummy classifier. Remarkably, structural features (45 UD in Table 1) performed better than lexical features (11 n-gram and 12 collgram in Table 1). One explanation could be the size of the corpus underlying the language models, which does not provide enough evidence for the frequency of items in non-translated Russian. Note that our lexical features do not directly rely on the frequencies of individual items. Instead, they estimate the ratios of high- and low-frequency items in translated and non-translated subcorpora of text chunks. Combining two lexical feature sets brings the performance of the classifiers to the area of 70-81% accuracy, with the exception of translations from French (66%). It seems that n-grams and collgrams complement each other in capturing lexical distinctions from the reference corpus. It is not surprising, given that frequencies of OOV n-grams and collgrams are commonly picked among the top five predictors in each feature set by most of the classifiers. The combined features accumulate the effects of the individual feature sets and are harder to interpret. They can be used to generally demonstrate the extent to which translations are distinguishable from non-translations. In our experiment, the accuracies of translations from any SL on all features were in the range from 82% to 91%.

To explore the impact of the individual features on the classification outcome, we employed Recursive Feature Elimination (RFE) algorithm based on Support Vector Regressor (SVR), which selected a unique combination of N features. RFE selects features that return the best classification results. There were only two features shared in RFE-based selections by all classifiers, if N=30: the ratio of high frequency trigrams and the frequency of modal predicates (defined as the the cumulative frequency of lemma мочь, lemma следовать with a dependent infinitive, three modal adverbs (можно, нельзя, надо) and 11 adjectives in the short form from a modal predicative list (e.g. должен, способный, возможный)). However, the frequency analysis for these features shows no significant differences in their values in many test subcorpora.

Even if the difference was significant at $p < 0.05$, the effect size did not exceed Cohen's d of 0.2.

RQ2 Interestingly, the results from the binary classifications for lexical features, unlike UD features, are more volatile with regard to the assumption that translations from more distant languages are more predictable. On structural features, this statement holds with the bold exceptions of Finnish and Bulgarian which returned too low or too high classification results for 84-91% and 79-83% brackets for distant and close languages, respectively. However, on either of the lexical feature sets, there is hardly any consistency to be seen.

In feature selection, there were no features that cut across all language pairs among top N=10 and N=20 UD features. However, typologically similar languages shared up to five features. For example, the selections for English, German and Swedish included 'ccomp', 'mquantif', 'xcomp', 'nnargs', 'ppron'[3]. We interpret this finding as evidence of the SL traces in translations, leaving a more in-depth analysis of these features for future work.

To find out whether the SL signal in each language pair is strong enough to make the respective translations stand out of the bulk of other translations and non-translations, we applied a multiclass classification on the entire feature set. In this experiment, each of the 12 subcorpora (Russian reference and 11 source-language subcorpora) was classified against the rest of the subcorpora. We achieved the overall accuracy of 53% (F-measure of 0.53), which was well above the random baseline of 6%. The darker areas in the confusion matrix (Figure 2 in Appendix) show that translations from same-family SLs were confused more often between themselves than with translations from distant languages. For instance, Belarusian is often confused with Ukrainian, but never with English and Spanish. In a similar manner, Russian non-translations were more likely to be misclassified as translations from Ukrainian, Belarusian or Latvian than from English or Swedish.

The results for individual languages also confirm the hypothesis that SL footprints specific for each language group are discernible in translations. Particularly, distant SLs generate translations that are more distinguishable as such than translations from closer languages. Better results (darker blue squares in Figure 2) are achieved for more distant languages, e.g. French, English and Swedish (F-measure of 0.62, 0.58 and 0.58, respectively). At the same time, the results are worse for typologically closer languages, such as Bulgarian and Latvian (both with an F-measure of 0.46). Against our expectations, the scores for the closest languages, such as Ukrainian and Belarusian, are not the lowest (0.50 and 0.54, respectively). This may be explained by fewer instances that we have for the lowest-scoring Bulgarian and Latvian.

[3]The description of these features is taken from (Kunilovskaya and Lapshinova-Koltunski, 2020) and is omitted here due to space considerations.

5 Conclusion

We analysed translationese in literary texts, exploring differences between fiction originally authored in Russian and fiction translated into Russian from 11 languages. We found out that overall, we can automatically predict translations in this register with the accuracy well above the chance level. Besides, we compared performance of classifiers for translations from various SLs. We expected that translations from more distant SLs would return higher results in a series of binary classifications and would be easier to recognise in a multiclass setting than translations from typologically closer languages. This was confirmed for the subset of structural features, but not for lexical features, which returned mixed results. Possible explanations for the opposite tendencies observed in the data when using lexical features could be deficiency of the reference model and variation in literature style or literary translation tradition.

At the same time, we also understand the limitations of our data selection and research design. For example, experiments demonstrate that average sentence length might have been a better indicator of genre comparability than the length of a literary piece. The reference model for collocational features should probably be trained on a larger corpus to ensure greater coverage of lexical items. We leave detailed statistical analysis of the best-performing features for each language group and case studies based on parallel concordances for future work.

References

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.

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.

Yves Bestgen and Sylviane Granger. 2014. Quantifying the development of phraseological competence in l2 english writing: An automated approach. *Journal of Second Language Writing*, 26:28–41.

Johannes Bjerva, Robert Östling, Maria Han Veiga, Jörg Tiedemann, and Isabelle Augenstein. 2019. What do language representations really represent? *Computational Linguistics*, 45(2):381–389.

Gerlof Bouma. 2009. Normalized (Pointwise) Mutual Information in Collocation Extraction. In *Proceedings of GSCL*, pages 31–40.

Dmitrij Dobrovol'skij and Ludmila Pöppel. 2017. Constructions in parallel corpora: A quantitative approach. In *International Conference on Computational and Corpus-Based Phraseology*, pages 41–53. Springer.

Koel Dutta Chowdhury, Cristina España i Bonet, and Josef van Genabith. 2020. Understanding Translationese in Multi-view Embedding Spaces. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6056–6062, Barcelona, Spain. Online.

Stefan Evert and Stella Neumann. 2017. The impact of translation direction on characteristics of translated texts : A multivariate analysis for English and German. In Gert De Sutter, Marie-Aude Lefer, and Isabelle Delaere, editors, *Empirical Translation Studies: New Methodological and Theoretical Traditions*, volume 300 of *TILSM series*, pages 47–80. Mouton de Gruyter.

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.

Stefan Th Gries. 2010. Useful statistics for corpus linguistics. *A mosaic of corpus linguistics: Selected approaches*, 66:269–291.

Kenneth Heafield. 2011. KenLM: Faster and Smaller Language Model Queries. In *Proceedings of the EMNLP 2011 Sixth Workshop on Statistical Machine Translation*, pages 187–197. Association for Computational Linguistics.

Leena Kolehmainen and Helka Riionheimo. 2016. Literary Translation as Language Contact: A Pilot Study on the Finnish Passive. *International Journal of Literary Linguistics*, 5(3):1–32.

Yekaterina Krasnopeyeva. 2016. Russian translated discourse research: patterns of lexis usage as linguistic indicator of translation universals representation. *International Journal of Russian Studies*, 5(1):80–92.

Maria Kunilovskaya. 2017. Linguistic tendencies in English to Russian translation: the case of connectives. In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2017"*, volume 2, pages 221–233.

Maria Kunilovskaya and Ekaterina Lapshinova-Koltunski. 2020. Lexicogrammatic Translationese across Two Targets and Competence Levels. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4102–4112, Marseille, France. The European Language Resources Association (ELRA).

Gerard Lynch and Carl Vogel. 2012. Towards the automatic detection of the source language of a literary translation. In *Proceedings of COLING 2012*, pages 775–784, Mumbai. Association for Computational Linguistics.

Dmitry Nikolaev, Taelin Karidi, Neta Kenneth, Veronika Mitnik, Lilja Saeboe, and Omri Abend. 2020. Morphosyntactic Predictability of Translationese. *Linguistics Vanguard*, 6(1).

Maeve Olohan. 2001. Spelling out the optionals in translation: a corpus study. *UCREL Technical Papers*, 13:423–432.

Marius Popescu. 2011. Studying translationese at the character level. In *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP 2011)*, September, pages 634–639. Association for Computational Linguistics.

Tiina Puurtinen. 2003. Genre-specific Features of Translationese? Linguistic Differences between Translated and Non-translated Finnish Children's Literature. *Literary and Linguistic Computing*, 18(4):389–406.

Ella Rabinovich, Noam Ordan, and Shuly Wintner. 2017. Found in Translation: Reconstructing Phylogenetic Language Trees from Translations. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 530–540, Vancouver, Canada. Association for Computational Linguistics.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. http://is.muni.cz/publication/884893/en.

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.

Maurizio Serva and Filippo Petroni. 2008. Indo-European languages tree by Levenshtein distance. *EPL (Europhysics Letters)*, 81(3).

Ilia Sominsky and Shuly Wintner. 2019. Automatic Detection of Translation Direction. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1131–1140, Varna, Bulgaria. INCOMA Ltd.

Lucia Specia, Gustavo Henrique Paetzold, and Carolina Scarton. 2015. Multi-level translation quality prediction with QUEST++. In *ACL-IJCNLP 2015 - 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, Proceedings of System Demonstrations*, pages 115–120. Association for Computational Linguistics.

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

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

A Appendix

type	number	list of features
morphological forms	8	past tense, passive voice form, finite and two non-finite forms of verb (infinitive and all participles), deverbal nouns, superlative and comparative degrees of comparison
word classes	7	personal, indefinite, possessive and demonstrative pronouns, adverbial quantifiers, coordinate and subordinate conjunctions
discourse markers	6	contrastive, additive, causative-consecutive, temporal-sequential connectives, epistemic stance markers and 'but' as a separate feature.
types of clauses	7	relative clauses, pied-piped subtype of relative clauses, correlative constructions, adverbial clauses introduced by a pronominal adverb, adjectival clause, clausal complement, a predicative or clausal complement without its own subject
syntactic features	8	words in attributive function, modal predicates, auxiliary, passive voice auxiliary, copula verbs, nouns or proper names used in the functions of core verbal argument and as subject of a passive transformation, asyndeton
general text measures	9	lexical type-to-token ratio and lexical density (based on disambiguated content types), mean hierarchical and mean dependency distances, number of simple sentences, negative sentences, interrogative sentences, number of clauses per sentence, sentence length

Table 2: UD features: features extracted from the UD-annotated documents.

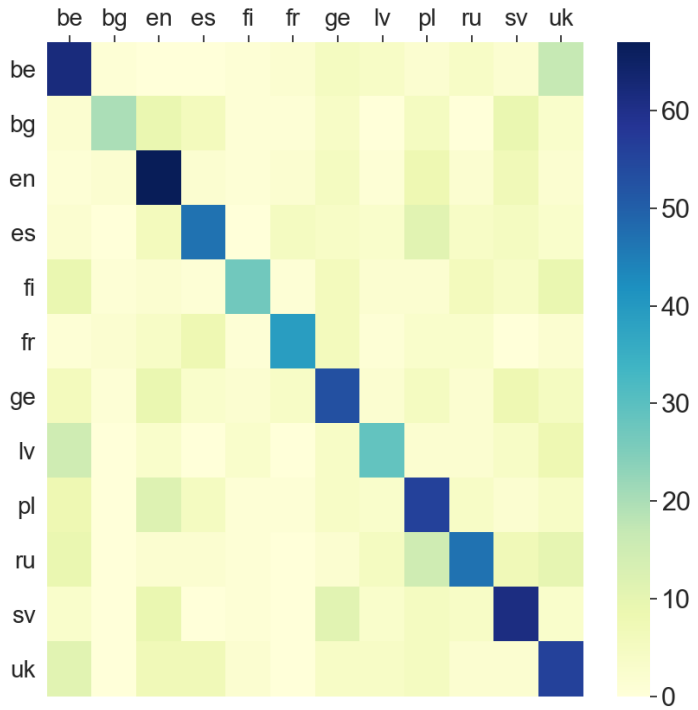

Figure 2: Confusion metrics for multiclass classification (the colour captures the number of predicted datapoints).