Translationese in Russian Literary Texts

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

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

Abstract

The paper reports the results of a translationese study of literary texts based on translated and non-translated Russian. We aim to find out if translations deviate from non-translated literary texts, and if the established differences can be attributed to typological relations between source and target languages. We expect that literary translations from typologically distant languages should exhibit more translationese, and the fingerprints of individual source languages (and their families) are traceable in translations. We explore linguistic properties that distinguish non-translated Russian literature from translations into Russian. Our results show that non-translated fiction is different from translations to the degree that these two language varieties can be automatically classified. As expected, language typology is reflected in translations of literary texts. We identified features that point to linguistic specificity of Russian non-translated literature and to shining-through effects. Some of translationese features cut across all language pairs, while others are characteristic of literary translations from languages belonging to specific language families.

1 Introduction

In this paper, we pursue a comparative analysis of literary texts that include texts originally authored in Russian and Russian translations from a variety of source languages (SLs) using machine learning methods.

We seek to establish if and how non-translated literature in Russian differs from literary translations. More specifically, we test if literary translations from typologically distant languages (in relation to the target language [TL]) exhibit more translationese, and hence, are easier to predict in a text classification scenario.

There are only a few computational studies based on literary translations. Intuitively, literary texts should defy generalisation as each author's style and each literary work is expected to manifest unique forms of artistic expression. Besides, literary translation is viewed as an independent creative endeavour, which, under the dominant covert translation strategy, results in highly readable and fluent texts. Logically, translations in this register might be less dependent on their source texts in terms of frequency of linguistic features. To the best of our knowledge, there were no large-scale investigations into Russian translations of literary texts.

With a view to analysing translationese in literary texts, we formulate two research questions.

First, this study seeks to establish if non-translated literary texts differ from literary translations, and if they do, which linguistic properties distinguish non-translated Russian literature from translations into Russian (RQ1).

Second, this study has the objective to find out if the fingerprints of individual SLs and their typological families can be traced in Russian translations (RQ2).

It has been shown that translationese "dialects" generated by typologically similar languages tend to demonstrate more similarities than those coming from typologically distant SLs (see Section 2). To explore the impact of SLs on the target text properties, we use a set of machine learning experiments automatically classifying translated and non-translated fiction in Russian in binary and multiclass scenarios. This study reports the classification performance and identifies discriminative features of translated and non-translated literature. We conjecture that this investigation can yield empirical evidence of various translational strategies pursued in individual language pairs. All linguistic inconsistencies in relations between translations and non-translations for specific language pairs that we identify can be further tackled from the point of view of critical translation studies, which focuses on social and cultural practices in translation.

Proceedings of LaTeCH-CLfL 2021, pages 101–112
Punta Cana, Dominican Republic (Online), November 11, 2021.

2 Related Work

Translationese and literary texts Translationese studies analyse linguistic features shared by translations that make them distinct from comparable non-translated texts (Gellerstam, 1986; Baker, 1993). The deviations of translations from comparable originally-authored texts in the target language (referred to as 'non-translations' in this paper) are revealed through the analysis of translationese indicators, i.e. linguistic features that exhibit a different frequency distribution in translations as compared to non-translations. These features are used to automatically distinguish translations from non-translations (Baroni and Bernardini, 2006; Volansky et al., 2015; Rabinovich and Wintner, 2015; Rubino et al., 2016; Kunilovskaya and Lapshinova-Koltunski, 2020). The choice of candidate translationese indicators is usually motivated by expectations of specific translational behaviours based on cross-linguistic contrastive properties of a given language pair or by (theoretically) known patterns in translations. Over the decades of translationese research, several dozens of features were tested, with some being especially effective in a number of settings. Features used for analysis of translationese are also widely used in register analysis, variational linguistics and stylometry. Several studies have shown that combinations perform better than individual features in machine learning setting for a number of translationese-related tasks (Lynch and Vogel, 2012; Evert and Neumann, 2017; Sominsky and Wintner, 2019). However, it is also a fact that translationese indicators can be specific for language pairs and registers, and literary translation is an Most corpus-based translationese studies of literary texts offer a single-feature analysis and do not employ machine learning (ML), e.g. passive constructions (Kolehmainen and Riionheimo, 2016), *that*-complementiser (Olohan, 2001), phrasal verbs (Cappelle and Loock, 2017), complex non-finite constructions, clause connectives and keywords (Puurtinen, 2003).

Popescu (2011) and Lynch and Vogel (2012) are the only works known to us that use ML approach to the study of translationese in literary texts. The first study compared originally-written English literary texts with translations from French and German using character 5-grams. The second study reports the results of SL detection based on English literary translations from Russian, German and French. Their four-class classifier achieved the accuracy of 80% on a combined set of 50 features, including such surface features as frequencies of most-used part-of-speech (PoS) bigrams, 15 top token unigrams (mostly functional words) and 19 document-level features such as sentence- and word-length and readability scores. The authors analysed some of the top-ranking features trying to link them to the SLs. We propose a similar experiment: we run a multiclass classification on a set of features trying to identify best predictors. However, we rely on very different – more interpretable, language-specific and translationally-motivated – features and experiment on a twice bigger data from a wider range of SLs. The mapping of individual features to known tendencies in translational behaviour was shown to be disputable, with one feature being associated with various tendencies (Zanettin, 2013; Kunilovskaya and Corpas Pastor, 2021), and we refrain from establishing the indicator-tendency link in the top-down manner.

Source language detection Source language detection task utilises the fact that translations tend to retain enough traces of the SLs involved (shining-through effect, Teich, 2003) for the benefits of linguistic typology and practical purposes of establishing translation direction in parallel corpora. For instance, Rabinovich et al. (2017) showed that the signal from the source language is enough to restore the phylogenetic languages trees. The authors experimented on English translations of European parliamentary speeches from languages representing three families and showed that classification errors most frequently occurred between languages from the same family. SLs with isomorphic structures tend to share more features in translations to a third language. That idea was exploited by Bjerva et al. (2019) who explored genetic, geographical and structural distances between languages and found that similarities between embeddings learnt from translations correlated best with the structural distance values obtained from vectorised treebanks of respective SLs meaning that translations carried over structural properties of their sources, while geographic coordinates and genetic distances calculated from phylogenetic trees did not correlate with the properties of translations as well. Dutta Chowdhury et al. (2020) showed that the more isomorphism was detected between translations into English and non-translated English, the closer the source languages were to English. They

learned delexicalised multi-view representations – embeddings based on PoS, semantic tags, and synsets – and used isomorphism between embedding spaces. However, all these studies used parliamentary speeches, which belong to a more conventionalised text type than fiction. These texts, as well as their translations, are more homogeneous in style in comparison to literary texts used in the present paper. The translators are also more constrained in their linguistic choices by the requirements of official and documentary nature of texts. All the studies above cover English as the target language, whereas we focus on Russian.

A recent study by Hu and Kübler (2021) showed that translationese features not discriminative for English translations worked for a different language. The authors analysed Chinese translations of journalistic texts from seven SLs in a ML classification approach. This study stands out, because it defined Chinese-specific translationese features, while most research in the field relied on easily-extractable language-independent features. They made observations on typological traces of the SLs manifested in the TL-unusual frequencies of certain feature groups.

SL-TL distance and the strength of the SL signal in translations is also key to solving the task of translation direction detection: Sominsky and Wintner (2019) found that the more distant were the source and the target languages, the higher the results. The authors achieved accuracy of 80-90%, even though they were challenged to perform classification at sentence-level. Their feature set was designed to address the sparsity of data, inevitable if typical document-level features were used. They relied on alignment of phrases identified as 'minimal translation units' and represented with PoS. The relation between predictability of translations and the divergence between the source and the target languages was also analysed by Nikolaev et al. (2020). The authors used entropy measures to check whether morphosyntactic entropies of original-language corpora are significantly different from those of corpora containing translations[1]. The results showed that translations from similar and distant languages were predictable 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 translation specificity. In translations from highly-divergent languages, however, translators tended to produce non-idiomatic translations, that were not recognised by models trained on the target language.

[1]The authors use a small pre-existing Parallel Universal Dependencies corpus of 1000 sentences from the news domain and Wikipedia translated from English into eight languages

3 Research Design

3.1 Data

In the current work, all data comes from the parallel subcorpus of translations into Russian of the Russian National Corpus (RNC). We followed a rigid data selection procedure to reduce possible confounding factors. Our sampling frame is aimed at reducing the influence of idiolects, while giving us enough data per language pair for machine learning experiments. It includes the following criteria, applied recursively:

- size and number of available documents (at least 30,000 tokens);
- unique combination of author and translator to avoid the influence of the over-represented authors or translators;
- no auto-translations or novels by bilingual authors (e.g. Nabokov, Vasil Bykov);
- translations produced after 1940 till now (sources are more widely spread in time).

The resulting sample consists of 11 translational subcorpora. We chunked this corpus to ensure direct comparability of results between language pairs and to properly balance our data. To this end, we randomly selected six books from each subcorpus, making sure that the average sentence length of the selected books is in the empirically-established optimal range (9, 22) to provide for generic homogeneity of the data (this constraint excluded novels with considerable parts written in verse or as drama).

From each of the selected books, we randomly extracted 15 chunks of about 150 consecutive sentences. We preferred chunking by the number of sentences rather than by word count as many of our features are normalised to the number of sentences. A comprehensive overview of the size of data in each subcorpus after chunking and lemmatisation is given in Figure 1. Translational data used in our experiments roughly totals 2,538,951 tokens (168,683 sentences), with each subcorpus represented by 90 chunks totalling at least 200,000 tokens (with

the exception of Belarusian (187,000 tokens and Swedish 189,000 tokens), the number of sentences being roughly 15,000 per source language. The SLs in the resulting corpus come from four language families: Romance (French, Spanish), Germanic (Swedish, English, German), Balto-Slavic (Baltic: Latvian, Slavic: Polish, Belarusian, Ukrainian, Bulgarian) and Uralic (Finnish) based on phylogenetic languages tree in (Serva and Petroni, 2008). Figure 1 also shows a subcorpus of non-translations (marked as Russian). It was built following the sampling and chunking principles described above from the monolingual part of the RNC.

All translational subcorpora are in one-sentence-per-line format. We discarded by-lines and headings, such as "Chapter 5", "Jane Eyre" and "Charlotte Bronte", empty lines and lines without alpha, including cases where the absence of text was marked with "—". The 11 translational subcorpora and non-translations were annotated within Universal Dependencies (UD) framework using UDPipe (v1, Straka and Straková, 2017).[2] The chunks in the resulting conllu format were used as input to our feature extraction module.

3.2 Features

Fist and foremost, this research aims to shed a light on how the Russian language of translated fiction is different from that used in the original Russian prose. This motivates the selection of the research design focused on features. We included a few well-known translationese indicators and a number of Russian-specific features extracted from parsed sentences. Such features as sentence length, content-lemma-based type-to-token ratio, lexical density, frequencies of discourse markers are known to predict translations well in a range of target languages and registers.

The motivation for including primarily structural features is manifold: (i) these features are less vulnerable to sparsity given a chunk size of around 2000 tokens (used e.g. in Volansky et al., 2015; Hu and Kübler, 2021); (ii) it is a common practice in translationese studies to refrain from using lexical features to avoid the impact of domain differences between translations and non-translations; and (iii) various morphosyntactic features (frequency of demonstratives, relative clauses, modal predicates, etc.) were shown to be effective in translationese detection. In a previous study based on Russian, these structural features were shown to work much better than lexical features (such as n-gram ranks and PMI scores, see Kunilovskaya and Corpas Pastor, 2021). Of several dozens of UD relations, we use the subset of seven relations that was shown to perform well for translationese detection in English-to-Russian mass-media texts (Kunilovskaya and Kutuzov, 2018).

We also included features that are susceptible to change in translation, according to textbooks on practical translations. Due to contrastive differences, translations into Russian (from Germanic languages, at least) are expected to feature higher frequencies of:

- pronominal determiners, e.g. этот, тот, весь, каждый, некоторый (this, that, all, every, some);
- possessive pronouns, e.g. мой, твой, его (my, your, his);
- relative and adverbial clauses that might be used for unpacking dense and unusual syntactic constructions in other languages;
- modal predicates – Он может знать ответ (He might know the answer) instead of a more typical Возможно, он знает ответ (Probably he knows the answer).

At the same time, we expect lower frequencies for:

- negative particles or main sentence negation,
- deverbal nouns and
- simple sentences.

Most of our extraction rules rely on the UD output, and we are constrained by the quality of annotation in this respect. Overall, our feature set includes 45 features[3]; their values are normalised frequencies of various UD tags and their combinations. The features include morphological forms (past tense, passive voice form, etc.), syntactic features (e.g. number of clauses per sentence, sentence length), word classes (e.g. types of pronominal function words, adverbial quantifiers), and dependency relations (e.g. adjectival clause, clausal complement). Besides, we include two features reflecting sentence complexity: mean hierarchical

[2] With the Russian model trained on SynTagRus treebank (v2.5, Droganova et al., 2018).

[3] We provide the full list of features in Appendix A.

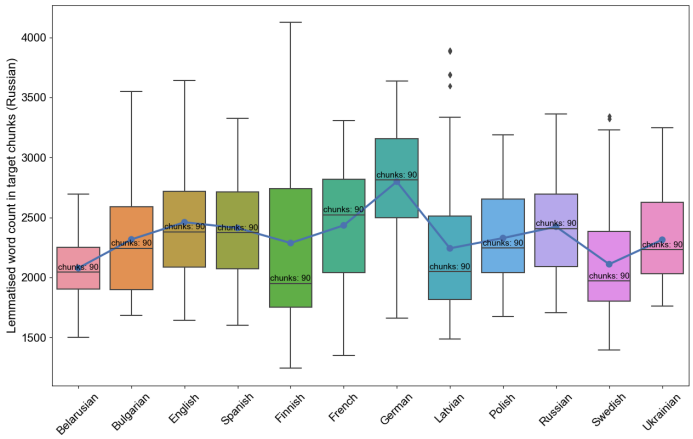

Figure 1: Distribution of chunk sizes (in tokens) in each subcorpus after lemmatisation

distance and mean dependency distance (described in Jing and Liu, 2015) and two features reflecting richness of vocabulary - type-to-token ratio and lexical density.

The normalisation basis varies depending on the type of feature as recommended in (Evert and Neumann, 2017): total tokens for word classes; number of sentences for conjunctions and modal predicates; total verbs for verb forms; total number of dependencies in the sentence for the selected types of dependencies.

3.3 Methods

We run two types of experiments: (i) a set of 11 binary classifications where we compare Russian translations from each of the SLs and fiction originally authored in Russian; and (ii) 12-class classification where we test whether data from any subcorpora stand out as very different from the rest of translated and non-translated Russian. Support Vector Machine (SVM) with Radial Basis Function (rbf) kernel is used as the main algorithm in both settings. Each feature was centred around the mean and scaled to feature standard deviation independently. The models are evaluated using cross-validation.

To avoid overfitting, we used scikit-learn *GroupKFold* algorithm which generated cross-validation folds where training and test sets in each split did not include chunks from the same books. In other words, in each iteration we trained the model on 10 books and tested on the two unseen books, one from each class. If this precaution is not in place, we may end up achieving unrealistically high accuracy for all binary classifications. While we have 6 books for each subcorpus, the maximum number of cross-validation splits that we can afford is six (book-aware 6-fold cross-validation). In the multiclass scenario, we used standard 10-fold cross-validation instead.

The multiclass classifier predicts the source language of translations or Russian for non-translations. In this experiment, we use one-vs-rest decision-making strategy, i.e a binary classifier is fitted on data splits, which cast each class against all the other classes. At test time, an unseen sample is predicted by all classifiers, and the most confident prediction is output.

For all ML experiments below we report accuracy and macro F1-score. Note that our 12 subcorpora are well-balanced with regard to the number of observations: translations from each SL and non-translations are represented by 90 chunks of text. A fair random chance level for binary classifiers is 50%, F1 = 0.5. For 12-class classification, a pseudo-classifier with random predictions on a uniform distribution achieves the accuracy of 9.8%, with F1-score 0.10.

To detect features particularly useful in our translationese classifications, we used two approaches: (i) ANOVA-based feature selection; and (ii) feature weights analysis. The first method utilises significant differences in feature values and helps us to find out an optimal subset of features for better classification results, while getting rid of possible noise in the data. We compare lists of selected features across all 11 classifications to see whether those involving translations from similar languages have a bigger intersection of best translationese indicators.

In the second approach, we followed Teich et al. (2016) and Argamon et al. (2008) to tap into feature weights returned by each classifier. Note, that to get access to feature weights we had to switch to

an SVM with a linear kernel[4]. The feature weights indicate which of the features from the list push the classifiers decision towards this or that class. The features assigned to a certain class are interpreted as distinctive for this particular class. The features with the greatest weights, averaged across all folds, are particularly important. Using this methodology, we aim to discover features that are associated with Russian non-translated literature as opposed to any translations in all 11 classification tasks. Similarly, the intersection of the features distinctive for translations in each of the 11 classifications can be seen as universal indicators of translationese in Russian translated literature.

4 Results and Discussion

4.1 Classifier performance

Binary scenario To answer the questions if Russian non-translated literary texts are distinct from literary translations, and if the scale of the differences can be linked to the SL typology, we compare the results from 11 binary classifications, where non-translated Russian is classified against translations from each of the 11 languages in our data (Russian vs. translations from English, vs. translations from Ukrainian, etc.). The results in Table 1 are arranged with regard to proximity to Russian – (1) Germanic, Romance and Uralic, i.e. typologically distant languages (in relation to Russian) including English (en), German (ge), Swedish (sv), Spanish (es), French (fr) and Finnish (fi), and (2) Balto-Slavic, i.e. typologically close languages, i.e. Latvian (lv), Belarusian (be), Ukrainian (uk), Polish (pl), Bulgarian (bg). Our hypothesis was that the more distant a language is with regard to Russian, the better the classification results.

One immediate observation from Table 1 is that all classification results were above the chance level of 50% accuracy. Translations from typologically distant languages returned higher accuracy (over 70% accuracy) than translations from Slavic languages (in the range from 59% to 67%). The two bold exceptions from this generalisation were translations from French and Bulgarian. Texts translated from Bulgarian were easier to distinguish from non-translated Russian literature than texts from other Slavic languages (accuracy 72.22%) This result could be attributed to the peculiar traits of the Bulgarian grammar that is different from other Slavic languages (e.g. absence of noun cases, infinitives, complex conjugation system). Novels translated from French were comparatively difficult to distinguish from non-translated Russian literature (accuracy 60.56%). A very similar low classification accuracy was demonstrated by translations from Ukrainian (59.44%). The greatest differences between translations and non-translations were seen for the German subcorpus, for which the accuracy reached 84.44%. The results for Slavic languages are consistently more than 2% lower than the accuracy for distant languages.

	language	accuracy	F-score
	en	75.56	0.74
	ge	*84.44*	*0.84*
	sv	71.11	0.70
(1)	es	74.44	0.74
	fr	**60.56**	**0.56**
	fi	70.00	0.67
	lv	61.11	0.59
	be	65.56	0.63
(2)	*uk*	*59.44*	*0.56*
	pl	67.22	0.64
	bg	**72.22**	**0.70**

Table 1: Results of 11 binary classifiers

It is noteworthy that relatively low translationese classification results as compared to results reported in other studies on translationese in fiction (Lynch and Vogel, 2012; Popescu, 2011) can be attributed to the rigour of our evaluation setup. If we used standard 10-fold cross-validation ignorant of the book-associated groups in the data, the binary classification results (with the same hyperparameter settings C=10.0, gamma=0.01 for the RGF kernel) would range from 89.44% (Belarusian) to 98.89% (Swedish). This is an indirect evidence that each translated book has its own unique and learnable structural peculiarities.

With regard to the features that might be less useful, we established that reducing the feature set by a third, to 30 features selected by ANOVA by each binary classifier independently, degraded the results by only 1% to 7% (for Finnish and Swedish there was no change in performance). On the reduced feature set the observations about the impact of typological differences on translationese properties were still standing. The description of the selected and discarded features appears in Section 4.2.

[4]The linear kernel accuracy for binary setting differs from the rbf kernel in the range from -10% for German to +5% for Ukrainian

Classification errors in multiclass scenario To further explore the differences in translationese dialects observed in translations from typologically-grouped SLs, we tried to distinguish the 12 classes representing SLs and Russian in our data in a multiclass scenario. In particular, we are interested in classification errors obtained from a confusion matrix that helped us to see if typologically close languages were more frequently confused with each other than more distant languages.

Overall, the 12-class classification achieved the accuracy of 65.4% with a macro F1-score of 0.68 (rbf kernel, 10-fold cross-validation, one-vs-rest strategy, C=0.1, gamma=0.01). The best classification results based on macro F1-score were seen for Russian (F1 = 0.78). Translations from Finnish were second most recognisable (F1 = 0.74). However, the Russian class had much higher recall (as expected, some translations, except those from German and Swedish, were erroneously classified as Russian non-translation). Note that true Russian texts were relatively rare mistaken for translations. The specificity of translations from Finnish was such that it did not attract many misclassifications. To visualise misclassification patterns and to see which SL translations were most often confused with each other, we drew a directed graph based on error statistics, see Figure 2. It has classes as nodes and the number of false positive errors for each language pair on edges. To distil the important error patterns, we retained only the edges with over 5 errors (5% of the size of the expected class).

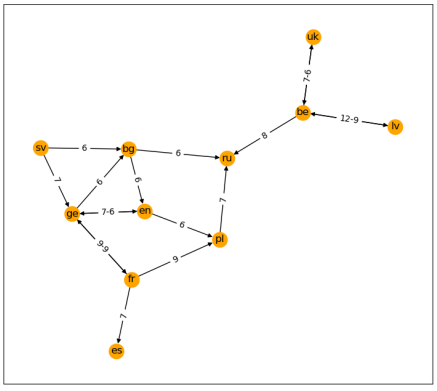

Figure 2: Errors in multi-class classification

Figure 2 confirms our expectations that translations from typologically similar languages are more often confused between themselves. The upper right corner of the graph has East Slavic languages and Latvian, which have the highest number of mutual misclassifications (esp. between Belarusian and Latvian) and are often confused with Russian. The lower part of the graph shows errors between Romance languages, while Germanic languages seem to be grouped together by error statistics on the left. Belarusian, Bulgarian and Polish are the three Slavic languages that are confused with non-translated Russian fiction. At the same time, both Bulgarian and Polish were also confused with either translations from English and other Germanic languages, or French, respectively. The error patterns that emphasised language families got stronger if we reduced the number of features to top 30 (not shown for considerations of space). One curious and unexpected observation from the error network analysis is that Bulgarian and Polish attracted more errors from non-Slavic languages than any other Slavic language.

4.2 Feature Analysis

Feature selection These experiments are aimed at finding the best-performing subset of features for all classifiers, if possible, and for individual SLs.

We assume that the intersection of the best features selected by each of the 11 binary classifiers contains translationese indicators that cut across all language pairs. We found that this intersection included only three features if we capped the number of best features (selected based on analysis of variance [ANOVA] results) at 30: contrastive connectives (`advers`), adverbial clause introduced by a pronominal (`whconj`) and demonstrative pronouns (`demdets`). However, their importance varied across the classifiers. This is confirmed by the difference in their ranks on the ordered lists of selected features and by the volatility of classifiers' performance on this subset of features. The lack of wide feature intersection in the best features for all classifiers indicates that translations from each SL have unique properties that distinguish them from non-translations, and using any universal subset of features can be sub-optimal to achieve the best classification results. To test this assumption, we found the number of features (in the range of 1 to 45) which returned the highest classification accuracy for each of the 11 classifications. Note that the performance of all classifiers in this setting was much higher than reported in Table 1 and ranged between 73.89 (Finnish) and 87.78 (German) for distant languages, and between 73.33 (Belarusian) and 82.78 (Bulgarian) for Slavic languages. It turned out that some translations were recognised best on just a

few features, while using more features degraded classification quality. For example, translations from Spanish, French and Polish were best separated from non-translations using just one feature (translations had significantly higher frequencies of relative clauses, demonstrative pronouns and clausal modifiers of nouns, respectively); translations from Swedish seem to have two strong translationese indicators: more demonstrative pronouns and fewer interrogative sentences.

Not surprisingly, close Slavic languages (Belarusian, Ukrainian) require larger feature sets (15 and 21 features, respectively) to achieve best classification results. Their distinctions from non-translations are more subtle and cannot be reliably captured by a few features. Even in this best setting, we were able to achieve 73.33% and 77.22% accuracy for translations from Belarusian and Ukrainian respectively.

At the same time, the largest feature set was called for to achieve the best accuracy for German: the classifier used 38 features to achieve the accuracy of 87.78%, which was higher than for any other SL. The second best result was returned for English: accuracy 86.67% on 14 features. The same feature set was particularly useful for capturing translationese in Russian and German translations from English (Kunilovskaya and Lapshinova-Koltunski, 2020), however on other text types than literature.

Every of the 45 features was selected by at least one of the 11 classifiers as part of the best-performing feature subset. Admittedly, it is best to develop language-pair-specific features for best results in each language pair. Using a universal set of count-based features in a setting similar to ours can introduce noise and degrade the quality of classification. However, a comprehensive feature set can be useful for exploratory purposes in frameworks where the metadata about SLs is not available. The most popular feature was demonstrative pronouns selected by eight classifiers, followed by lexical density, mean hierarchical distance, adverbial clauses and parataxis, found in five best feature sets. The least relevant features, appearing on just one of the best feature lists, included passive voice auxiliaries, lexical type-to-token ratio, copula verbs, temporal-sequential connectives and clausal complements.

With regard to the features selected for typologically similar SLs, English and German rely on 13 shared features (of 14 features selected to identify English translations). The best classification setups to distinguish translations from Belarusian and Ukrainian from Russian non-translations share six features. Interestingly, English-German and Belarusian-Ukrainian intersection had only one feature in common (adverbial clause introduced by a pronominal – `whconj`). This supports the hypothesis that translations from similar SLs have similar translationese properties.

Features of non-translated Russian The analysis of feature weights from the binary classifiers yielded five features associated with the non-translations class in all 11 binary classifications: `simple` (simple sentences), `interrog` (interrogative sentences), `mpred` (modal predicates), `pasttense` (past tense), `ccomp` (clausal complements). This means that these five features were characteristic for Russian literary texts if compared to literary translations into Russian from any language. However, their weights varied across the 11 classification tasks. None of them was shared in the 10 most weighted features, indicating that the classifier found more useful (more weighted) features to take the decision in each given subcorpus.

Past tense was selected among the top translationese predictors in all classifications except the one against translations from Finnish, while interrogative sentences (i.e. ending in '?') were important for Russian in all classifications except those against translations from Ukrainian, Latvian and Finnish.

Interestingly, sentence length was a strong indicator of Russian class and appeared in top 10 features in 9 classifications. It means that using feature weights and ANOVA for identifying distinctive features returned intersecting, but not identical results, which should be taken into consideration in further literary analysis.

A glance at the data table reveals that interrogative sentences were more frequent in Russian non-translated texts than in translations, whereas modal predicates had lower frequency. The low frequency of modal predicates in non-translated literature is in accord with our assumptions about the differences between translated and non-translated Russian (see Section 3.2 above). Interrogatives are typical of narrative texts containing dialogues between characters, but they can also be used as rhetorical means to actively engage the reader as in example (1).

(1) Тут меня охватило раздражение – <кого-кого>? В <каком> округе? Да <что> за китайская грамота такая? (Then I was seized by irritation – who? Which county? What double Dutch is that?)

High frequency of interrogatives may also indicate the authority relations between interactants (Halliday and Matthiessen, 2014).

In our chunks, in most source text collections there were fewer interrogative sentences than in Russian non-translations. Russian translations actually introduced additional question marks, a typical normalisation behaviour. Can it be a sign of the specificity of Russian fiction compared to other national literatures? However, a deep qualitative analysis is needed to find a true explanation of this observation.

Source language translationese Using the same technique as applied for the Russian non-translations, we searched for the overlaps between most-weighted features associated with the translations class in 11 binary classifications. We found five overlapping features, universal for all literary translations in our data: `lexdens` (lexical density), `demdets` (demonstrative determiners), `relativ` (relative clauses), `advers` (adversative discourse markers) and `whconj` (adverbial clauses introduced by a pronominal adverb). However, similarly with the results observed for the non-translations, none of them appears in all lists of top 10 predictors of the translated class. At the same time, `lexdens` has high weights associated with translations from 9 languages (all except German and French), `demdets` – with 8, `relative` – with 7, `advers` and `whconj` – both with 6.

Remarkably, all of them show higher frequencies in translated Russian (with an exception of lexical density). This result is in line with expectations for the four features, based on what is known about translational behaviour from earlier studies: `demdets, relativ, whconj, advers`. For example, the increased use of the adversative markers (наоборот, несмотря на, однако) expressing the relation of contrast and comparison, could indicate explicitation in translation, as there are studies showing that this type of relation is cognitively more complex, and thus is more frequently expressed with explicit signals than other discourse relations (Hoek et al., 2017).

The observation on the low value of lexical density (i.e ratio of PoS disambiguated content words types to the total number of running words) in non-translated Russian if compared to translations into Russian deserves a comment. Since translated texts are predominantly reported to have a lower lexical density than non-translations (see e.g. Laviosa, 1998; Steiner, 2012), our findings are surprising. Steiner (2012) reports lower lexical density for fiction translated from German into English than for the English non-translated texts in the same register. At the same time, this tendency may differ for translations of fiction. According to register studies (see Biber et al., 1999, p. 62), fictional texts are characterised by lower lexical density values (if compared to general language) due to dialogues and complex purposes of fictional writing that combines informational aspects and aesthetic concerns. This means that translations in our data may show a higher degree of colloquial, spoken or source-culture-specific elements than non-translated Russian literature, which relies on a narrower range of vocabulary or has a higher proportion of non-content words.

5 Conclusion

We focused on the phenomenon of translationese in Russian literature to find out that Russian non-translated literary texts are rarely confused with translations, and thus are very distinct. Translated Russian literary texts differ from non-translated ones to the degree that they can be automatically detected and the source language signal is strong enough to be traced in translations. As assumed, typologically close languages are more frequently confused with each other than more distant languages, which points to language typology being reflected also in translations of literary texts. The analysed features that distinguish translations and non-translations help to uncover specificity of Russian non-translated literature. We were also able to detect some universal features of translationese in fiction, as well as language-specific and language-typology-specific translation features. We plan to use these results as guidance in a more qualitative analysis of the features that behave differently in translated language, and of language pairs (esp. Bulgarian and French) which were found diverting from the observed trend, in the framework of critical translation studies.

References

Shlomo Argamon, Jeff Dodick, and Paul Chase. 2008. Language use reflects scientific methodology: A corpus-based study of peer-reviewed journal articles. *Scientometrics*, 75(2):203–238.

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.

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.

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.

Bert Cappelle and Rudy Loock. 2017. Typological differences shining through : The case of phrasal verbs in translated English. In Gert De Sutter, Marie-Aude Lefer, and Isabelle Delaere, editors, *Empirical Translation Studies. New Theoretical and Methodological Traditions*, pages 235–264. Mouton de Gruyter.

Kira Droganova, Olga Lyashevskaya, and Daniel Zeman. 2018. Data conversion and consistency of monolingual corpora: Russian ud treebanks. In *Proceedings of the 17th international workshop on treebanks and linguistic theories (tlt 2018)*, volume 155, pages 53–66.

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. *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.

MAK Halliday and Ruqaiya Hasan. 1976. *Cohesion in English*. Equinox.

Michael A.K. Halliday and Christian M.I.M Matthiessen. 2014. *Halliday's introduction to Functional Grammar. Revised by Christian M.I.M. Matthiessen*. Routledge, London.

J. Hoek, S. Zufferey, J. Evers-Vermeul, and T.J.M. Sanders. 2017. Cognitive complexity and the linguistic marking of coherence relations: A parallel corpus study. *Journal of Pragmatics*, 121:113–131.

Hai Hu and Sandra Kübler. 2021. Investigating translated chinese and its variants using machine learning. *Natural Language Engineering*, 27(3):339–372.

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.

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.

Maria Kunilovskaya and Gloria Corpas Pastor. 2021. Translationese and register variation in English-to-Russian professional translation. In *New Perspectives on Corpus Translation Studies*. Springer.

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 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)*, Marseille, France. The European Language Resources Association (ELRA).

Sara Laviosa. 1998. Core patterns of lexical use in a comparable corpus of English narrative prose. *Meta*, 43:557–570.

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.

Ella Rabinovich and Shuly Wintner. 2015. Unsupervised identification of translationese. *Transactions of the Association for Computational Linguistics*, 3:419–432.

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.

Erich Steiner. 2012. A characterization of the resource based on shallow statistics. In Silvia Hansen-Schirra, Stella Neumann, and Erich Steiner, editors, *Cross-Linguistic Corpora for the Study of Translations: Insights from the Language Pair English-German*, volume 11 of *Text, Translation, Computational Processing [TTCP]*. Walter de Gruyter GmbH, Berlin/Boston.

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.

Elke Teich, Stefania Degaetano-Ortlieb, Peter Fankhauser, Hannah Kermes, and Ekaterina Lapshinova-Koltunski. 2016. The Linguistic Construal of Disciplinarity: A Data Mining Approach Using Register Features. *Journal of the Association for Information Science and Technology (JASIST)*, 67(7):1668–1678.

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

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

A Appendix: Features

- 8 morphological forms: degrees of comparison (`comp, sup`), passive voice (`shortpassive, bypassive`), non-finite forms of verb (`infs, pverbals`), nominalisations (`deverbals`) and finite verbs (`finites`);
- 7 word classes: pronominal function words (`ppron, demdets, possdet, indef`), adverbial quantifiers (`mquantif`), coordinate and subordinate conjunctions (`cconj, sconj`);
- 7 UD relations following (Kunilovskaya and Kutuzov, 2018): 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`);
- 4 syntactic functions: various PoS in attributive function (`attrib`), modal predicates (`mpred`), copula verbs (`copula`), nouns or proper names used in the functions of core verbal arguments(`nnargs`);
- 7 syntactic features for sentence type and structure: simple sentences (`simple`), number of clauses per sentence (`numcls`), negative sentences (`neg`), types of clauses – relative (`relativ`) and pied-piped subtype (`pied`), correlative constructions (`correl`), adverbial clause introduced by a pronominal ADV(`whconj`);
- 2 graph-based features: mean hierarchical distance and mean dependency distance (`mhd, mdd`) (Jing and Liu, 2015);
- 5 list-based features for semantic types of discourse markers (`addit, advers, caus, tempseq, epist`) and *but* (not followed by 'also' and not in the absolute sentence end). (`but`). The semantic classification roughly follows (Halliday and Hasan, 1976; Biber et al., 1999; Fraser, 2006);
- 2 text measures: lexical variety, i.e. ratio of PoS-disambiguated content word types to their tokens (`lexTTR`) and lexical density, i.e. ratio of PoS-disambiguated content words types to all tokens (`lexdens`). PoS-disambiguation is based on 'lempos' annotation ((look_VERB vs look_NOUN)); content PoS include ADJ, ADV, VERB, NOUN.